MLA

Software for MultiLevel Analysis of Data with Two Levels

User's Guide for Version 1.0b

Frank M. T. A. Busing [1]
Erik Meijer [2]
Rien van der Leeden [3]

December, 1994

[1]Frank M. T. A. Busing is Research Associate in the Department of Psychometrics and Research Methodology at Leiden University, P.O. Box 9555, 2300 RB Leiden, The Netherlands.

[2]Erik Meijer is Graduate Student in the Department of Psychometrics and Research Methodology at Leiden University.

[3]Rien van der Leeden is Assistant Professor in the Department of Psychometrics and Research Methodology at Leiden University

Preface

This manual describes `MLA`, Version 1.0b, a computer program developed for multilevel analysis of data with two levels. The `MLA` program can be characterized by four major properties:

- User-friendly interface.
- Extensive options for simulation, in particular, three options for bootstrapping multilevel models.
- Simple estimation methods, providing an alternative for the complex iterative estimation procedures that are commonly used to estimate the parameters of multilevel models.
- A fast algorithm, using the Broyden-Fletcher-Goldfarb-Shanno optimization method to obtain maximum likelihood estimates of all model parameters.

The `MLA` program runs as a stand-alone batch program on 286-, 386- and 486-based personal computers under `DOS`. It uses simple `ASCII` text files as input and output. The program is easy to use by means of a number of statements starting with a keyword. Models are specified by simply formulating the model equations.

This manual provides the necessary information for the new user to fit multilevel models with two levels to a hierarchical data set. It is expected that the user has basic knowledge of regression analysis. A brief introduction to multilevel analysis and related concepts is given in the first chapter. References to three major textbooks on multilevel analysis can be found in the text.

The `MLA` program was developed, and is being further developed, by Frank Busing, Erik Meijer, and Rien van der Leeden. As the version number indicates, this manual describes a beta version. We are still doing research to polish and improve certain simulation options. We would very much appreciate hearing about any of your experiences using the program and this manual. Please contact us by email:
`busing@rulfsw.leidenuniv.nl` or `vanderleeden@rulfsw.leidenuniv.nl`.

We would like to thank Jan de Leeuw and Ita Kreft for helpful discussions, comments, and references.

The Institute for Educational Research in the Netherlands (SVO) is gratefully acknowledged for supporting this project by a grant (SVO, project no. 93713).

Frank M. T. A. Busing
Erik Meijer
Rien van der Leeden

Leiden, December 1994

Contents

1 Introduction **1**
- 1.1 Introduction to multilevel analysis . . . 1
- 1.2 Why another program for multilevel analysis? . . . 5

2 Theory **9**
- 2.1 The general two-level model . . . 9
- 2.2 Descriptive statistics . . . 11
- 2.3 Ordinary Least Squares . . . 12
- 2.4 Maximum Likelihood methods . . . 14
- 2.5 Residuals . . . 14
- 2.6 Posterior means . . . 15
- 2.7 Diagnostics . . . 16
- 2.8 Simulation . . . 16
- 2.9 Missing data . . . 24

3 Input **27**
- 3.1 `/TITLE` (optional) . . . 27
- 3.2 `/DATA` (required) . . . 28
- 3.3 `/MODEL` (required) . . . 29
- 3.4 `/CONSTRAINTS` (optional) . . . 30
- 3.5 `/SIMULATION` (optional) . . . 30
- 3.6 `/TECHNICAL` (optional) . . . 33
- 3.7 `/OUTPUT` (optional) . . . 35

4 Output **37**
- 4.1 Analysis of variance . . . 37
- 4.2 Analysis of covariance . . . 41
- 4.3 Repeated measures analysis . . . 43
- 4.4 Multilevel analysis . . . 45
- 4.5 Simulation study . . . 47

A Technical Appendix **49**
- A.1 The model and the likelihood function . . . 49
- A.2 Some useful formulas . . . 50
- A.3 Computational formulas for the function and gradient . . . 54
- A.4 The asymptotic covariance matrix of the estimators . . . 58
- A.5 Reparametrization . . . 68

B Read.Me **71**

References **75**

Chapter 1

Introduction

1.1 Introduction to multilevel analysis

Multilevel analysis comprises a set of techniques that explicitly take into account the hierarchical structure in the data. In this section, a brief introduction to the underlying ideas of multilevel analysis is given. Several relevant topics, such as hierachical data structures, intra-class correlation, the formulation of a multilevel model, and the estimation of the model parameters are discussed. This introduction does not contain formulas. Chapter 2 will discuss the main formulas, and the Technical Appendix will give supplementary mathematical details.

Hierarchical data

Hierarchically structured data arise in a variety of research areas. Such data are characterized by so-called "nested" membership relations among the units of observation. Classical examples of hierarchically structured data are found in educational research where, for instance, students are nested within classes and classes are nested within schools. But, in many other instances in the social and behavioral sciences, as well as in many other fields of science, data are also hierarchically structured. For instance, in clinical psychology, clients can be nested within therapy groups, people can be nested within families, and so forth. A sociological example is given by a study concerning employees nested within industries.

It should be noted that nested structures naturally arise where explicit hierarchical sampling schemes are used. This is often the case in large scale educational research where, for instance, a set of schools is sampled first, followed by the sampling of a set of students within these schools. However, there are many other cases where data are not explicitly sampled in that way, but where it appears to be a fruitful approach to treat them as having a hierarchical structure. For instance, in a medical study one could consider it to be important that patients can be viewed as nested within general practitioners. Apart from this, there are several types of data for which it proves to be very useful to apply the concept of hierarchy, because it makes their analysis more easy and transparent. One example is the hierarchical treatment of repeated measures data, where measurements at different points in time are considered nested within individuals. Another example is the analysis of data from meta analysis, where, say, p-values can be treated as being nested within studies, providing a (partial) solution for the problem of comparing apples with oranges.

With hierarchical data, it is common to have information obtained from the different

levels in the hierarchy. For instance, one has variables describing the individual students, but also variables describing their schools. When analyzing such data one has to decide in what way the hierarchical structure of the data is taken into account. Obviously, the easiest approach is simply ignoring the structure and analyzing the data at the student level, leaving all school information for what it is. Generally, however, one's intention will be to use all information in the data, and use it correctly. Thus, if one is also interested in school differences and in their possible interaction with effects measured at the student level, one has to solve the "unit-of-analysis" problem. This means that one has to decide whether to analyze the data at the student level, incorporating disaggregated variables from the school level, or to analyze the data at the school level, incorporating aggregated variables from the student level. Unfortunately, according to De Leeuw in his introduction to the book of Bryk and Raudenbush (1992), both of these strategies are subject to serious disadvantages. Hence, traditional "single level" analyses fail in the presence of nested data.

Intra-class dependency

The basic problem with hierarchical data is that group membership may cause *intra-class dependency*: People from the same group are more alike than people from different groups. The reason for this phenomenon is that people within a group share the same environment, have the same leader, experiences, and so forth. In other words, people within the same group have the same score on a number of variables, most of them never measured and thus omitted from any possible model. Hence, if we fit a (common, single level) model to such data, intra-class dependency has an effect on the error terms. It causes the error terms to be correlated. The result is that the usual assumption of independent observations is violated if the nested structure of the data is ignored. The degree of intra-class dependency is reflected in the *intra-class correlation*. Obviously, this idea of intra-class dependency applies to every hierarchical data set. Their intra-class correlations, however, may differ substantially.

Multilevel models

For the analysis of hierarchical data, hierarchical models, or "multilevel" models have been developed. Such models can be conceived as linear regression models, specified separately for each level of the hierarchy, but statistically connected. Since each level of the hierarchy has its own regression equation, predictor variables measured at either level can be included in the appropriate level model.

Because hierarchical data structures frequently arise in social and behavioral science research, but also in many other scientific areas, the application and development of multilevel analysis has in the last decade drawn a lot of attention from numerous researchers. Below, a brief introduction of some relevant topics concerning multilevel models will be given. A more comprehensive introduction of these topics is given by Kreft and Van Der Leeden (1994). For extensive discussions on theory and application of multilevel analysis, we refer to the textbooks by Goldstein (1987), Bryk and Raudenbush (1992), and Longford (1993).

Small example

A small, imaginary, example from education may clarify what is meant by a multilevel model. Suppose we have data of students nested within schools, and we want to predict

the score on a math test from the amount of time spend on doing math homework. Furthermore, we expect smaller schools to be more effective than larger ones, so we collect the school size as another variable. Clearly, at the student level, 'math' is the dependent variable, and 'homework' is the predictor variable. At the school level, 'size' is the predictor variable. Now the multilevel model for this example, in this case a two-level model, is specified as follows. At the student level, Level-1, for each school a regression model is formulated with 'math' as the dependent variable and 'homework' as the predictor. This reflects the intra-class dependency of the observations (the students) within each school: All models contain the same variables, but we expect them to yield different intercept and slope estimates within each school. At the school level, Level-2, a regression model is formulated in which the intercepts and slopes of the first-level models are dependent variables, predicted by the second-level variable 'size'. This reflects the possible effect of school size on school effectiveness: School size may influence the estimated relationship between 'math' and 'homework'.

At first glance, the model presented above seems to lead to a hierarchically structured regression procedure, which proceeds in two steps: First, the models for all schools are estimated, and then the intercept and slope estimates are used as the dependent variables in the Level-2 model, which is then estimated. Although such procedures have been proposed in the past, this is not what will be discussed here under the heading of multilevel models, because there is no statistical connection between the Level-1 and Level-2 models. In multilevel models, separate regression equations for each level are only formulated because they facilitate insight and understanding. The statistical linkage of both levels is created by the Level-2 model which states that Level-1 regression coefficients—intercepts and slopes—are treated as *random variables* at the second level. The Level-2 model models intercept and slope estimates as a mean value over all schools plus a school-specific deviation or residual. It follows that we are not primarily looking for intercept and slope estimates for each separate school, but for their means and variances (and their covariance) over all schools. In this way, just as students are considered a sample from a population of students, schools are considered a sample from a population of schools.

There are several reasons why it may be useful to consider the school-specific coefficients as random. First, the schools in the data set are usually a random sample from the "population" of schools, and scientists are usually interested in the population, rather than the specific data set. Second, with a model that explains part of the variation in the random coefficients, the effect of the school-level variables on the student-level relationships can be assessed, and, in particular, the model can give guidance to schools that want to improve their effectiveness. Third, the relationships between the outcome variable and the student-level predictors become clearer: Between-school variation that may blur these relationships is accounted for, and consequently, the estimates of the average coefficients are more precise.

School-specific estimates of intercept and slope can, however, be obtained. This will be discussed below under the heading of Random Level-1 coefficients.

Cross-level interaction

If a school-level predictor variable like 'size' is added to the Level-2 model in our imaginary example, means and variances change to conditional means and variances. It means that part of the variance of intercepts and slopes among schools is explained by 'size'. The contribution of this school-level variable introduces a term to the model that specifies a relationship between both levels: The relationship between 'size' at the school level and the slope coefficient for each separate school, which is part of the model at the student level.

As was said above, this term refers to the expected influence of 'size' on the regression of 'math' on 'homework'. In the terminology of multilevel analysis this term is called a *cross-level interaction*. For some researchers, this interaction term provides the main attraction to multilevel analysis. It is the cross-level interaction parameter that leads to the interpretation of "slopes-as-outcomes" (cf. Aitkin & Longford, 1986).

The number of levels

Theoretically, we can model as many levels as we know the hierarchy has, or as we think it will have. In practice, however, most applications of multilevel analysis concern problems with two or three levels. Data sets with more than three levels are rare. In fact, a majority of applications just concerns two-level data and can be viewed as "within-and-between-analysis" problems. It should be noted that models with more than three levels show a rapid increase in complexity, especially where interpretation is concerned. If such models are necessary, they should be limited to rather simple cases, that is, to cases with only a few predictor variables.

Random Level-1 coefficients

In multilevel modeling, we are usually not looking for estimates of the regression coefficients within each separate group, but for their variances and covariances. However, there can be circumstances in which we still want to obtain the "best" estimates for these coefficients, also called *random Level-1 coefficients*. Such questions may arise, for example, in education when schools are to be ranked in terms of effectiveness, using their estimated slope coefficients (Kreft & De Leeuw, 1991). The first thing that comes to mind is to simply estimate them by a separate (`OLS`) regression for each school. However, this procedure has the serious disadvantage that the coefficients will not be estimated with the same precision for each school. For instance, in one school, we could have, say, 45 students, whereas in another school we only have 7 students. This will definitely influence the accuracy of results.

Within the framework of multilevel analysis there is a way to obtain best estimates of these coefficients by a method called *shrinkage* estimation. The underlying idea of this estimation is that there are basically two sources of information: the estimates from each group separately and the estimates that could be obtained from the total sample, ignoring any grouping. Shrinkage estimation consists of a weigted combination of these two sources. The more reliable the estimates are within the separate groups, the more weight is put on them. Vice versa, the less reliable these estimates are, that is, the less precise, the more weight is put on the estimates obtained from the total sample. The result is that estimates are "shrunken" towards the mean of the estimates over all groups. The amount of shrinkage depends on the reliability of the estimates from the separate groups. The less precise the estimates are, the more they are "shrunken" towards the mean over all groups.

Technically, the shrunken estimators are the expectations of the (random) coefficients given the parameter estimates and the data of all groups.

Estimation

Fitting a multilevel model amounts to fitting one combined model, instead of separate models for each level. It is the translation of the idea that, although separate models for each level may be formulated, they are statistically connected, as was mentioned in a

previous subsection. The combined model contains all relevant parameters. In the next chapter, we will further clarify this subject.

Combined models, or multilevel models, can be viewed as special cases of the general mixed linear model (cf. Harville, 1977). Such models are characterized by a set of fixed and a set of random regression coefficients. The parameters that have to be estimated are the fixed coefficients and the variances and covariances of the random coefficients and random error terms. The fixed coefficients are informally called *fixed parameters* and the variances and covariances of the random coefficients and random error terms are informally called *random parameters*, although all these parameters are technically nonrandom. They are the parameters associated with the fixed and random parts of the model, respectively.

To obtain estimates for the parameters, several estimation procedures have been proposed. These procedures are all versions, in one way or another, of full information (`FIML`) or restricted maximum likelihood (`REML`). `FIML` and `REML` estimators have several attractive properties, such as consistency and efficiency. A drawback of both approaches, however, is their relative complexity. Generally, parameter estimates must be obtained iteratively and serious computational difficulties may arise during such processes.

Software

The flourishing of models and techniques for analyzing hierarchical data has been stimulated by the software widely available for estimating multilevel models. The three major packages are `ML3` (Prosser, Rasbash, & Goldstein, 1991), `VARCL` (Longford, 1990) and `HLM` (Bryk, Raudenbush, Seltzer, & Congdon, 1988), although multilevel models can also be estimated with `BMDP` (`BMDP-5V` procedure, Schluchter, 1988), `SAS` (`MIXED` procedure, SAS Institute, 1992), and `GENMOD` (a program based on the work of Mason, Wong, & Entwistle, 1983). The three major packages use different methods for maximizing the likelihood. In `ML3` an *Iterative Generalized Least Squares* (`IGLS`) procedure is implemented (Goldstein, 1986), and a restricted version of `IGLS` (`RIGLS`; Goldstein, 1989). `VARCL` uses Fisher scoring (Longford, 1987) and `HLM` uses the `EM` algorithm (Dempster, Laird, & Rubin, 1977; Bryk & Raudenbush, 1992). A comparative study of several of these programs is given in Kreft, De Leeuw, and Van Der Leeden (1994).

Final remark

In the literature multilevel models are referred to under various names. One may find the terms *random coefficient regression models* (De Leeuw & Kreft, 1986; Prosser et al., 1991), *contextual effects models* (Blalock, 1984), *multilevel mixed effects models* (Goldstein, 1986), *random parameter models* (Aitkin & Longford, 1986), *full contextual models* (Kreft & Van Der Leeden, 1994), *variance components models* (Aitkin & Longford, 1986), *multilevel linear models* (Goldstein, 1987; Mason et al., 1983), and *hierarchical linear models* (Bryk & Raudenbush, 1992).

Although there are minor differences, all these models are basically the same. In one way or another they are versions of the multilevel model discussed here, or straightforward extensions thereof.

1.2 Why another program for multilevel analysis?

This manual describes the use and capabilities of a new program for multilevel analysis, called `MLA`. This program has been developed to analyze data with a two-level hierarchical

structure. In this section we will explain why we think it is useful to add a new program for multilevel analysis to the existing ones mentioned above. In other words, we are concerned with the question: What is special about MLA?

Simulation options

Much research concerning multilevel analysis has been directed to the extension and refinement of multilevel theory, including the development of multilevel software, and to applications in other domains than educational research. At the same time, however, several relevant questions of a statistical nature concerning this development are still not answered fully satisfactorily. One major problem is that estimates of parameters and standard errors, as well as hypothesis tests based on them, rely on large sample properties of the estimates. Unfortunately, little is known about the behavior of the estimates when sample size is small (Raudenbush, 1988). Goldstein (1987) even suggests optimizing the design of a multilevel study by the use of pilot or simulation studies. An additional problem is that it is usually assumed that the error terms are normally distributed. In practice, this assumption will often be violated, which has other undesirable consequences for using standard error estimates for hypothesis testing and construction of confidence intervals.

Fortunately, there is an increasing number of simulation studies available, which give insight into the quality of estimates of parameters and standard errors under various conditions (Busing, 1993; Van Der Leeden & Busing, 1994; Kreft, 1994). Concerning empirical data sets, however, we think that extensive simulation options, in particular options for bootstrapping, would be a very useful addition to a program for multilevel analysis.

Therefore, four different simulation methods are implemented in MLA:

1. A bootstrap method that uses the estimated parameters as "true values" of the parameters of a multivariate normal distribution from which new outcome variables are drawn. This method is implemented in the ML3 program as well (as far as we know this is the only multilevel analysis program so far that has some form of simulation option built in). It is called the *parametric bootstrap*.

2. A bootstrap method that uses the observed values of outcome and predictor variables for resampling. Thus, whole cases are resampled. Therefore, we call it *cases bootstrap*.

3. A bootstrap method that uses estimates of the error terms at both levels for resampling. In contrast with the cases bootstrap, this method leaves the regression design unaffected. We call it *error bootstrap*. Because the error terms at both levels must be estimated in order to be resampled, we need the estimates for the separate Level-1 models (random Level-1 coefficients). As was explained earlier, there are two choices for these coefficients: OLS estimates for each group separately, or shrinkage estimates, based on the whole sample. These two choices account for two additional options that can be used when applying the error bootstrap.

4. The *jackknife*. With this method, one entire case is deleted for each resample. There are as many resamples as there are cases.

Depending on the type of simulation used in MLA and depending on the nature of the data, the user can decide to resample both levels in the data, or only the first or the second level. This feature may be useful, for instance, in analyzing repeated measures data.

Alternative simple estimation methods

Usually, complicated iterative estimation procedures are used to estimate the parameters of multilevel models. From a theoretical and technical point of view, these procedures provide the best estimates that can be obtained. However, in practice, some of the algorithms used may be rather slow under certain conditions. In other cases serious computational difficulties may arise that are not easy to overcome. De Leeuw and Kreft (1993) discuss alternative estimation procedures for both fixed and random parameters in multilevel models that are non-iterative and relatively easy to implement. Moreover, in certain cases the quality of the parameter estimates is rather good. Hence, one could question the real gain of the complicated iterative procedures over these simpler alternatives. Therefore, in `MLA`, we have implemented a one-step and a two-step `OLS` procedure. A simple `WLS` procedure has still to be implemented.

Simple procedures can always be used as an addition to complex ones, and vice versa. Their results can always be compared with the results of the iterative methods. It depends on the data which estimation procedures are to be preferred (De Leeuw & Kreft, 1993; Kreft, 1994).

Fast Maximum Likelihood algorithm

To maximize the likelihood function, the `MLA` program uses the *Broyden-Fletcher-Goldfarb-Shanno* (`BFGS`) method (Press, Flannery, Teukolsky, & Vetterling, 1986). This is a fast and stable method to optimize arbitrary functions. It requires that the function and the gradient (the vector of first derivatives) of the function with respect to the parameters be programmed. It minimizes the function with respect to both fixed and random parameters simultaneously. As such, it resembles most the algorithm used by `VARCL`, although the `BFGS` method does not compute the inverse of the information matrix at each iteration. The algorithms of `ML3` and `HLM` alternately update the fixed and the random parameters.

Chapter 2

Theory

In this chapter, the theoretical background of the general two-level model will be discussed. It will give the relevant formulas of the model equations, and it will give the theory and formulas of the descriptive statistics and estimators that are implemented in the program. Additionally, it will discuss the theory of the residuals, the estimators of the group-specific coefficients, and the diagnostic statistics that the `MLA` program provides, and the simulation options that can be chosen.

2.1 The general two-level model

In `MLA`, the following general two-level model is implemented. Suppose data are obtained from N individuals nested within J groups, with group j containing N_j individuals. Now, for group j ($j = 1, \ldots, J$), y_j is a vector containing values on an outcome variable, X_j is an $N_j \times q$ matrix with fixed, explanatory variables (including the constant), β_j is a vector of regression coefficients, and ε_j is a vector with random error terms (vectors and matrices of appropriate dimensions). Then, for each group j, the Level-1 or within-group model can be written as

$$y_j = X_j \beta_j + \varepsilon_j. \tag{2.1}$$

The Level-2 or between-group model can be written as

$$\beta_j = W_j \gamma + u_j, \tag{2.2}$$

where W_j is a $q \times p$ matrix with explanatory variables (including the constant) obtained at the group level, γ is a vector containing fixed coefficients and u_j is a vector with error terms. Equation (2.2) clearly illustrates the "slopes-as-outcomes" interpretation, because it gives the illusion that the coefficients in β_j are outcome variables in a separate Level-2 model.

However, substition of Equation (2.2) into Equation (2.1) gives the "total" model equation

$$y_j = X_j W_j \gamma + X_j u_j + \varepsilon_j. \tag{2.3}$$

This is a *mixed linear model* (Harville, 1977) of the form

$$y_j = X_j^* \gamma + Z_j u_j + \varepsilon_j, \tag{2.4}$$

in which $X_j^* = X_j W_j$ and $Z_j = X_j$. Several authors use different notations for the models presented in this chapter and in subsequent chapters. We find the separate model

equations (2.1) and (2.2) for the two levels most useful for interpretation of the model and its estimates, and the program input is therefore based on them (see chapter 3). For theoretical purposes, we find the form (2.4) most useful, where usually X_j^* will be simply written as X_j. Therefore, in the following both representations will be used where appropriate, and it will be clear from the context which form is used. For now, we will proceed with the form (2.4).

Generally, it is assumed that $\varepsilon_j \sim N(0, \sigma_\varepsilon^2 I_{N_j})$ and $u_j \sim N(0, \Theta)$, where σ_ε^2, the variance of the Level-1 error term, is an unknown (scalar) parameter, and Θ, the covariance matrix of the Level-2 error terms, is a (symmetric) matrix of unknown parameters. The covariance matrix V_j of y_j conditional on X_j and Z_j, that is, the matrix containing the variances and covariances of the random part $Z_j u_j + \varepsilon_j$ in Equation (2.4) conditional on Z_j, is expressed as

$$V_j = Z_j \Theta Z_j' + \sigma_\varepsilon^2 I_{N_j}. \tag{2.5}$$

A model for the complete data follows straightforwardly from stacking the J groups' models in Equation (2.4). Its equation is

$$\begin{pmatrix} y_1 \\ \vdots \\ y_J \end{pmatrix} = \begin{pmatrix} X_1 \\ \vdots \\ X_J \end{pmatrix} \gamma + \begin{pmatrix} Z_1 & 0 & \cdots & 0 \\ 0 & Z_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & Z_J \end{pmatrix} \begin{pmatrix} u_1 \\ \vdots \\ u_J \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_J \end{pmatrix},$$

or,

$$y = X\gamma + Zu + \varepsilon. \tag{2.6}$$

The covariance matrix of the complete data, conditional on X and Z, is

$$\begin{aligned} V = & \begin{pmatrix} Z_1 & 0 & \cdots & 0 \\ 0 & Z_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & Z_J \end{pmatrix} \begin{pmatrix} \Theta & 0 & \cdots & 0 \\ 0 & \Theta & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Theta \end{pmatrix} \begin{pmatrix} Z_1' & 0 & \cdots & 0 \\ 0 & Z_2' & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & Z_J' \end{pmatrix} \\ & + \sigma_\varepsilon^2 I_N \\ = & \begin{pmatrix} V_1 & 0 & \cdots & 0 \\ 0 & V_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & V_J \end{pmatrix}. \end{aligned}$$

The parameters of the model that have to be estimated are the fixed coefficients (elements of the vector γ), the covariance matrix Θ of the random coefficients, and the variance σ_ε^2 of the errors. The elements of γ are called the *fixed parameters*, and σ_ε^2 and the elements of Θ are called the *random parameters*.

In the following, formulas are presented for the various parts of the output of `MLA`. The order of this chapter is similar to the order of the output of `MLA`, as will become clear later on. In section 2.2 the computational formulas are presented for the descriptive statistics. The next section, 2.3, discusses various forms of ordinary least squares estimation, namely `OLS` estimates for each group separately (section 2.3.1) and one-step and two-step `OLS` for the fixed and random parameters of the total two-level model (sections 2.3.2 and 2.3.3,

respectively). Maximum likelihood estimation is dealt with in the next section (2.4), subdivided into subsections about full information maximum likelihood (section 2.4.1) and restricted maximum likelihood (section 2.4.2). An extensive elaboration on the subjects of maximum likelihood estimation will follow in Appendix A. In section 2.5 several types of residuals will be discussed, namely total residuals, raw residuals and shrunken residuals. Section 2.6 will introduce the posterior means and section 2.7 will discuss diagnostics. The theory behind the simulation options in `MLA` is described in section 2.8. Finally, in section 2.9, some remarks will be made about missing data.

2.2 Descriptive statistics

`MLA` produces (if asked for) the following descriptive statistics: mean, standard deviation, variance, skewness, and kurtosis. Any statistical package will produce these statistics as well. Before looking at the other output, it may be useful to inspect these statistics. Their formulas are:

$$\begin{aligned}
&\text{mean:} && \widehat{\mu} = \frac{1}{N}\sum_{i=1}^{N} X_i, \\
&\text{standard deviation:} && \widehat{\sigma} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \widehat{\mu})^2}, \\
&\text{variance:} && \widehat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \widehat{\mu})^2, \\
&\text{skewness:} && \widehat{\mu}_3 = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{X_i - \widehat{\mu}}{\widehat{\sigma}}\right)^3, \\
&\text{kurtosis:} && \widehat{\mu}_4 = \left[\frac{1}{N}\sum_{i=1}^{N}\left(\frac{X_i - \widehat{\mu}}{\widehat{\sigma}}\right)^4\right] - 3,
\end{aligned}$$

where X_i is the measurement of individual i on a typical variable X and N is the total sample size.

Another descriptive statistic that is provided is the *Kolmogorov-Smirnov Z* statistic. This is a measure of deviation from the normal distribution. It tests whether the observed variable has a normal distribution. It is defined as the maximum distance between the estimated (empirical) cumulative distribution function and the best-fitting cumulative normal distribution function. It is computed as follows (Stephens, 1974). First, sort the values of a given variable X, such that X_1 is the smallest value and X_N is the largest. Then compute $w_i = (X_i - \widehat{\mu})/\widehat{\sigma}$, $i = 1, \ldots, N$, and $z_i = \Phi(w_i)$, where Φ is the cumulative distribution function of the standard normal distribution. Now, Kolmogorov-Smirnov's Z is defined as

$$Z = \max_{1 \le i \le N}\left(\max\left\{z_i - \frac{i-1}{N}, \frac{i}{N} - z_i\right\}\right).$$

The (asymptotic) distribution of Z was derived by Durbin (1973), but it is too complicated to be implemented in `MLA` (it requires numerical integration and Fourier transformation). Stephens (1974), however, provides a table of critical values of a transformed statistic, $(\sqrt{N} - 0.01 + 0.85/\sqrt{N})Z$, which can be used to obtain a range of probability levels (p-values) indicating the significance of the deviation from normality. In `MLA`, p-values

are reported that are based on the assumption that the normal distribution is completely specified beforehand (not estimated). This is not entirely correct, because μ and σ are estimated, but it is sufficient for descriptive (exploratory) purposes. The formula used (cf. Mood, Graybill, & Boes, 1974, p. 509) is

$$\Pr(Z) = 2\sum_{j=1}^{\infty}(-1)^{j-1}e^{-2j^2NZ^2}, \tag{2.7}$$

where the series is truncated after convergence of the sum.

2.3 Ordinary Least Squares

2.3.1 Within-group models

In this section, we will use the notation (2.1)–(2.2). Consider Equation (2.1). Ordinary least squares estimates for β_j are given by

$$\widehat{\beta}_j = (X_j'X_j)^{-1}X_j'y_j, \tag{2.8}$$

and the estimated standard errors of the elements of $\widehat{\beta}_j$ are the square roots of the diagonal elements of the covariance matrix given by

$$\widehat{\mathrm{cov}}(\widehat{\beta}_j) = \widehat{\sigma}_j^2(X_j'X_j)^{-1}, \tag{2.9}$$

where

$$\widehat{\sigma}_j^2 = \frac{1}{N_j - q}(y_j - X_j\widehat{\beta}_j)'(y_j - X_j\widehat{\beta}_j), \tag{2.10}$$

and q is the dimension of β_j.

2.3.2 One-step `OLS` (total model)

From Equation (2.6) the term $Zu+\varepsilon$ can be considered the random part of the equation. Taking the total residuals

$$r = Zu + \varepsilon, \tag{2.11}$$

leaves, after substitution,

$$y = X\gamma + r. \tag{2.12}$$

Now, γ can be estimated using ordinary least squares. Notice that grouping is ignored. Estimates for γ are given by

$$\widehat{\gamma} = (X'X)^{-1}X'y. \tag{2.13}$$

Using the estimated residuals $\widehat{r} = y - X\widehat{\gamma}$, the estimate of the variance of the elements of r can be obtained by

$$\widehat{\sigma}_r^2 = \frac{1}{N-p}\sum_{i=1}^{N}\widehat{r}_i^2, \tag{2.14}$$

where p is the dimension of γ. This estimate $\widehat{\sigma}_r^2$ is the one-step OLS estimate of the variance of the residuals. The usual standard errors for $\widehat{\gamma}$ and $\widehat{\sigma}_r^2$ are, respectively,

$$\widehat{\mathrm{se}}(\widehat{\gamma}_l) = \sqrt{[\widehat{\sigma}_r^2 (X'X)^{-1}]_{ll}}, \tag{2.15}$$

$$\widehat{\mathrm{se}}(\widehat{\sigma}_r^2) = \widehat{\sigma}_r^2 \sqrt{\frac{2}{N-p}}. \tag{2.16}$$

2.3.3 Two-step OLS (total model)

With the two-step OLS, the same estimates $\widehat{\gamma}$ are used as with the one-step OLS, see (2.13). The total residuals for every group j can be divided into a Level-2 and a Level-1 part. This was already done in Equation (2.11). Using ordinary least squares, estimates for the Level-2 random components, u, can be obtained by

$$\widehat{u}_j = (Z_j' Z_j)^{-1} Z_j' \widehat{r}_j. \tag{2.17}$$

The estimate for the covariance matrix Θ of u becomes

$$\widehat{\Theta} = \frac{1}{J} \sum_{j=1}^{J} \widehat{u}_j \widehat{u}_j'. \tag{2.18}$$

The estimated covariances of the elements of $\widehat{\Theta}$ can be obtained by (Anderson, 1958, p. 161)

$$\widehat{\mathrm{cov}}(\widehat{\Theta}_{kl}, \widehat{\Theta}_{mn}) = (\widehat{\Theta}_{km}\widehat{\Theta}_{ln} + \widehat{\Theta}_{kn}\widehat{\Theta}_{lm})/J. \tag{2.19}$$

Consequently, the estimated standard errors of the elements of $\widehat{\Theta}$ are given by

$$\widehat{\mathrm{se}}(\widehat{\Theta}_{kl}) = \sqrt{(\widehat{\Theta}_{kk}\widehat{\Theta}_{ll} + \widehat{\Theta}_{kl}^2)/J}. \tag{2.20}$$

By first computing the residuals $\widehat{\varepsilon}$,

$$\widehat{\varepsilon} = \widehat{r} - Z\widehat{u}, \tag{2.21}$$

the estimate for σ_ε^2 becomes

$$\widehat{\sigma}_\varepsilon^2 = \frac{1}{N} \sum_{i=1}^{N} \widehat{\varepsilon}_i^2. \tag{2.22}$$

This estimate $\widehat{\sigma}_\varepsilon^2$ is the two-step OLS estimate of the variance of the elements of ε. The estimated standard error for $\widehat{\sigma}_\varepsilon^2$ becomes, analogous to Equation (2.16),

$$\widehat{\mathrm{se}}(\widehat{\sigma}_\varepsilon^2) = \widehat{\sigma}_\varepsilon^2 \sqrt{\frac{2}{N-q}}. \tag{2.23}$$

All the estimators in this section are consistent if $J \longrightarrow \infty$ and $N_j \longrightarrow \infty$ for each j. Although this may be unrealistic, these estimators may be good initial estimators (starting values) for maximum likelihood estimators. In some cases, the differences between these estimators and the maximum likelihood estimators is small, and therefore, these estimators can be used as well (Kreft, 1994; Van Der Leeden & Busing, 1994).

2.4 Maximum Likelihood methods

2.4.1 Full Information Maximum Likelihood (FIML)

One of the most important parts of the program consists of the maximum likelihood estimation. This estimation method was chosen for its desirable properties, such as consistency and efficiency. In maximum likelihood estimation, given the observations, parameters are found that maximize the likelihood function (Mood et al., 1974). This is the same as minimizing the minus-log-likelihood function. Assuming normally distributed errors, the density of y_j, given X_j and Z_j, is

$$f(y_j|X_j,Z_j) = \frac{1}{(2\pi)^{N_j/2}(\det V_j)^{1/2}} e^{-\frac{1}{2}(y_j - X_j\gamma)'V_j^{-1}(y_j - X_j\gamma)},$$

so that the contribution of Level-2 unit j to the minus-log-likelihood function is

$$\begin{aligned} L_j &= -\log f(y_j|X_j,Z_j) \\ &= \frac{N_j}{2}\log(2\pi) + \frac{1}{2}\log\det V_j + \frac{1}{2}(y_j - X_j\gamma)'V_j^{-1}(y_j - X_j\gamma) \end{aligned}$$

and the minus-log-likelihood function for the whole sample is simply the sum of all Level-2 units j, $L = \sum_{j=1}^{J} L_j$. This is the function that has to be minimized with respect to the parameters to obtain maximum likelihood estimators. Specifically, it will produce a set of fixed parameter estimates, $\widehat{\gamma}$, and a set of random parameter estimates, $\widehat{\Theta}$ for the second level and $\widehat{\sigma}_\varepsilon^2$ for the first level. Details can be found in Appendix A.

The asymptotic covariance matrix of the estimators is derived from the matrix of second derivatives of L (the Hessian matrix). This covariance matrix is used for the standard errors for both fixed and random parameters.

For a detailed description of the derivations used and an extensive discussion of the computational formulas used in the program, Appendix A contains all information.

2.4.2 Restricted Maximum Likelihood (REML)

This is an important alternative estimation procedure, which will be implemented in subsequent versions of the program.

2.5 Residuals

The total residuals are given by Equation (2.12),

$$\widehat{r} = y - X\widehat{\gamma}.$$

In this equation, estimated residuals $\widehat{r}$ are based on the fixed parameter estimates $\widehat{\gamma}$ from the maximum likelihood estimation, although other estimates of γ could be used as well.

The raw residuals for the first level are taken from the within-group model (2.1),

$$\widehat{\varepsilon}_j = y_j - X_j\widehat{\beta}_j, \tag{2.24}$$

where the estimates $\widehat{\beta}_j$ are the OLS estimates from (2.8). Using the between-group model from Equation (2.2) and the OLS estimates from Equation (2.8), the Level-2 raw residuals are

$$\widehat{u}_j = \widehat{\beta}_j - W_j\widehat{\gamma}, \tag{2.25}$$

where $\widehat{\gamma}$ stems from Equation (2.13).

The Level-2 shrunken residuals are given by

$$\widehat{u}_j = (Z_j\widehat{\Theta})'[Z_j\widehat{\Theta}Z_j' + \widehat{\sigma}_\varepsilon^2 I_{N_j}]^{-1}\widehat{r}_j. \tag{2.26}$$

where $\widehat{r}_j$ contains the total full information maximum likelihood residuals for group j (i. e., $\widehat{r}_j = y_j - X_j\widehat{\gamma}$, where $\widehat{\gamma}$ is the `FIML` estimator of γ), and $\widehat{\Theta}$ and $\widehat{\sigma}_\varepsilon^2$ are the `FIML` estimators of Θ and σ_ε^2, respectively. The formula is computationally rather inefficient. Therefore, the following more efficient formulas will be used. One can write

$$(Z_j\widehat{\Theta})' = \widehat{\Theta}Z_j'$$

and taking

$$\begin{aligned} \widehat{V}_j &= Z_j\widehat{\Theta}Z_j + \widehat{\sigma}_\varepsilon^2 I_{N_j} && \text{(from (2.5))}, \\ Z_j'\widehat{V}_j^{-1} &= \widehat{\sigma}_\varepsilon^{-2}\widehat{G}_j^{-1}Z_j' && \text{(from (A.12))}, \end{aligned}$$

where

$$\widehat{G}_j = I_q + Z_j'Z_j\widehat{\Theta}/\widehat{\sigma}_\varepsilon^2, \tag{2.27}$$

then

$$\begin{aligned} \widehat{u}_j &= \widehat{\Theta}Z_j'\widehat{V}_j^{-1}(y_j - X_j\widehat{\gamma}_j) \\ &= \widehat{\sigma}_\varepsilon^{-2}\widehat{\Theta}\widehat{G}_j^{-1}(Z_j'y_j - Z_j'X_j\widehat{\gamma}_j) \end{aligned}$$

Finally, the shrunken residuals for Level-1 follow from (2.11),

$$\widehat{\varepsilon} = \widehat{r} - Z\widehat{u}. \tag{2.28}$$

2.6 Posterior means

The posterior means are the shrunken estimators of β_j. They are the expected values of the β_j, given the data and the maximum likelihood estimates of γ, Θ, and σ_ε^2. They are derived from the shrunken residuals and their formula is

$$\widehat{\beta}_j = W_j\widehat{\gamma} + \widehat{u}_j, \tag{2.29}$$

where $\widehat{\gamma}$ is the estimate obtained by full information maximum likelihood and $\widehat{u}_j$ is taken from (2.26). This can easily be shown to be equal to

$$\widehat{\beta}_j = (I_q - \widehat{\Lambda}_j)(W_j\widehat{\gamma}) + \widehat{\Lambda}_j\widehat{\beta}_j^{\texttt{OLS}}, \tag{2.30}$$

where $\widehat{\beta}_j^{\texttt{OLS}}$ is the within-group `OLS` estimator (2.8) of β_j, $\widehat{\Lambda}_j = \widehat{\Theta}X_j'\widehat{V}_j^{-1}X_j$, $\widehat{V}_j = \widehat{\sigma}_\varepsilon^2 I_{N_j} + X_j\widehat{\Theta}X_j'$, and the notation is of (2.1) and (2.2). Thus, from (2.30), $\widehat{\beta}_j$ can be seen as a matrix-weighted average of the within-group estimator $\widehat{\beta}_j^{\texttt{OLS}}$ and the estimated prior expectation $W_j\widehat{\gamma}$ of β_j, the former being unbiased and the latter being more efficiently estimated. The more efficient $\widehat{\beta}_j^{\texttt{OLS}}$ is estimated, the more (matrix) weight it gets, and the closer the posterior means are to the within-group estimates. $\widehat{\Lambda}_j$ can be called an estimated "reliability" matrix (cf. Bryk & Raudenbush, 1992, p. 43).

2.7 Diagnostics

Currently, the only option for diagnostics performed by `MLA` apart from the descriptive statistics, is outlier detection. Although the term outlier seems to be unambiguous, this is not completely true. An outlier is considered to be a deviant observation in the data and not a deviant residual after model estimation. However, a procedure fitting outliers in the data as residual outliers is considered to be a robust procedure. Outliers are in `MLA` detected using residuals. So, we expect `MLA` to be robust against data outliers and therefore look for residual outliers. More research in the field of robustness for multilevel models would be useful, however.

The detection of outliers differs for Level-1 and Level-2 outliers. For both levels, the shrunken residuals are considered. For the first level, the quotients

$$\frac{\widehat{\varepsilon_{ij}}}{\sqrt{\widehat{\sigma}^2(\widehat{\varepsilon})}} \tag{2.31}$$

are calculated, where

$$\widehat{\sigma}^2(\widehat{\varepsilon}) = \frac{1}{N}\sum_{i,j}\widehat{\varepsilon}_{ij}^2$$

is the variance of the Level-1 residuals. Residuals will be displayed whenever the quotient (2.31), when compared to a standard normal distribution, has a p-value less than some (possibly) user-specified value. The default value is 0.1.

For the Level-2 outliers, the Mahalanobis distances of the Level-2 residuals to their theoretical mean of zero are calculated by

$$M_j = u_j'\widehat{\Theta}^{-1}u_j.$$

Now, residuals are displayed for which M_j is larger than the critical value corresponding to a (possibly) user-specified p-value of a chi-square distribution with q degrees of freedom, where q is the dimension of u. This p-value is the same as for the Level-1 outliers.

2.8 Simulation

The maximum likelihood theory discussed so far is based on a few assumptions, the most important of which are:

- The model (i. e., the conditional expectation $X\gamma$ and covariance matrix V) is correctly specified. The standard errors, t-values, exceedance probabilities, and likelihood ratio tests were derived under the condition that at least the (most general) model that is being estimated is correct in the population.

- The Level-1 (ε) and Level-2 (u) random errors are normally distributed. The likelihood function was derived under this assumption, and therefore, the `FIML` estimators and the estimators of their standard errors depend on it.

- The sample size is large. More specifically, the properties of the maximum likelihood estimators, such as their consistency, their (asymptotic) efficiency, and their (asymptotic) normal distribution, as well as the formulas for their standard errors were derived under the assumption that the sample size goes to infinity ($N \longrightarrow \infty$).

In practice, these assumptions will not be completely satisfied. One can only hope that they are met approximately. To be able to get an indication of how severe the finite sample size and possible nonnormality influence the results, the MLA program offers simulation options. In this section, the theory underlying these simulation options will be described. This focus will be on the possible bias of the estimates and on the possibly incorrect standard errors. More subtle information can, however, be extracted from the program by using a file to write the simulation results to.

The *bias* of an estimator $\widehat{\theta}$ of some parameter θ is defined as the difference between the expected value of the estimator and the true value of the parameter. A desirable property of an estimator is *unbiasedness*, which means that its bias is zero. In the maximum likelihood theory discussed so far, however, it was only stated that the FIML estimators are *consistent*. This means that as the sample size gets larger, the mean of the estimator converges to the true parameter value and its variance decreases to zero. Informally speaking, the estimator comes closer to the true parameter value as sample size gets larger. This is, of course, a highly desirable property, but it does not ensure that the estimator is unbiased in finite samples. In fact, maximum likelihood estimators are in many models and situations biased in finite samples. For a general class of regression models including multilevel models, however, Magnus (1978) proved that the maximum likelihood estimators of the *fixed* regression coefficients are unbiased. On the other hand, Busing (1993) showed in a Monte Carlo simulation study that the maximum likelihood estimators of the *random* parameters in multilevel models are biased.

The standard errors of the maximum likelihood estimators that are reported by MLA are derived from asymptotic theory. This means that they are based on the idea that as the sample size goes to infinity, the distribution of the estimators will converge to a (multivariate) normal distribution with a certain covariance matrix (see Appendix A). The reported standard errors that are the square roots of the diagonal elements of this matrix. The exceedance probabilities of the according t-values that are reported are based on the approximation of the distribution of the estimators by the normal distribution. In finite samples, this approximation may not be very good. The true standard errors may be quite different from the reported ones based on asymptotic theory, and the distributions of the estimators may not be normal. In fact, Busing (1993) showed in his simulation study that the distributions of the random parameters can be severely skewed. As mentioned above, however, the focus is on the bias and the standard errors and not on the specific distribution.

2.8.1 The jackknife

The *jackknife* was introduced by Quenouille (1949, 1956) to estimate the bias of an estimator from one sample, and to correct for it. Tukey (1958) proposed an accompanying estimator for the variance of the estimator, and hence for the standard error.

The idea of the jackknife is as follows. Consider an independently and identically distributed sample of size N from some distribution and an estimator $\widehat{\theta}_N$ of a parameter θ obtained from this sample. Many estimators based on a sample of size N have a bias that can be written as

$$\text{bias}_N = E(\widehat{\theta}_N) - \theta = \frac{b_1}{N} + \frac{b_2}{N^2} + \ldots, \tag{2.32}$$

where $b_1, b_2, \ldots$, are constants that do not depend on N. Now consider removing a group of m observations from the sample and reestimating θ based on this sample of size $N - m$. The resulting estimator may be called $\widehat{\theta}_{N-m}$. The estimator $\widehat{\theta}_{N-m}$ is of the same sort

as $\widehat{\theta}_N$. The only difference is that it is based on a sample of size $N-m$ instead of N. Therefore, the bias formula (2.32) also holds for this estimator, with $N-m$ substituted for N, that is,

$$\text{bias}_{N-m} = E(\widehat{\theta}_{N-m}) - \theta = \frac{b_1}{N-m} + \frac{b_2}{(N-m)^2} + \ldots \tag{2.33}$$

Now, consider the difference between (2.33) and (2.32), given by

$$E(\widehat{\theta}_{N-m} - \widehat{\theta}_N) = b_1 \left[\frac{m}{N(N-m)}\right] + b_2 \left[\frac{m(2N-m)}{N^2(N-m)^2}\right] + \ldots,$$

From this equation, it can be seen that an estimate of the leading term of the bias of $\widehat{\theta}_N$ can be obtained from

$$\begin{aligned} \widehat{\text{bias}}_{N,m} &= \frac{N-m}{m}\left(\widehat{\theta}_{N-m} - \widehat{\theta}_N\right) \\ &= \left(\frac{N}{m} - 1\right)\left(\widehat{\theta}_{N-m} - \widehat{\theta}_N\right). \end{aligned} \tag{2.34}$$

Now, a bias-corrected estimator of the parameter $\widehat{\theta}$ is

$$\begin{aligned} \widehat{\theta}^J_{N,m} &= \widehat{\theta}_N - \widehat{\text{bias}}_{N,m} \\ &= \frac{N}{m}\widehat{\theta}_N - \left(\frac{N}{m} - 1\right)\widehat{\theta}_{N-m}. \end{aligned} \tag{2.35}$$

From (2.32) and (2.33), it is found that the bias of this estimator is

$$E(\widehat{\theta}^J_{N,m}) - \theta = -\frac{b_2}{N(N-m)} + \ldots,$$

which is of order $1/N^2$ if m is relatively small compared to N. This is a much smaller order than the bias of $\widehat{\theta}_N$, which is of order $1/N$.

The estimator $\widehat{\theta}_{N-m}$ was obtained by removing one group of size m from the sample. There are, however, many groups that can be used for this. Consider, for example, the case that $m = 1$. Then there are N groups of size 1 that could be removed, namely all N observations. Now, call the estimator $\widehat{\theta}_{N-1}$ obtained by removing observation i from the sample $\widehat{\theta}_{(i)}$. The corresponding estimator of the bias is called $\widehat{\text{bias}}_{(i)}$, and the corresponding bias-corrected estimator is called $\widehat{\theta}^J_{(i)}$. Now, a more precise estimator of the bias can be obtained by averaging the different estimators of the bias:

$$\widehat{\text{bias}}_J = \frac{1}{N}\sum_{i=1}^{N} \widehat{\text{bias}}_{(i)}. \tag{2.36}$$

The corresponding bias-corrected estimator of θ is

$$\begin{aligned} \widehat{\theta}_J &= \widehat{\theta}_N - \widehat{\text{bias}}_J \\ &= N\widehat{\theta}_N - (N-1)\widehat{\theta}_{(.)}, \end{aligned} \tag{2.37}$$

where $\widehat{\theta}_{(.)}$ is the average of the estimators $\widehat{\theta}_{(i)}$: $\widehat{\theta}_{(.)} = \sum_{i=1}^{N}\widehat{\theta}_{(i)}/N$. The estimator $\widehat{\theta}_J$ is called the (ungrouped) jackknife bias-corrected estimator of θ and $\widehat{\text{bias}}_J$ is called the (ungrouped) jackknife bias estimator.

Tukey (1958) proposed to use the estimators $\widehat{\theta}_{(i)}$ to obtain an estimator of the variance of the estimator $\widehat{\theta}_N$. Its formula is

$$\widehat{\sigma}_J^2 = \frac{N-1}{N} \sum_{i=1}^{N} \left(\widehat{\theta}_{(i)} - \widehat{\theta}_{(.)} \right)^2 . \tag{2.38}$$

Although it was originally an estimator of the variance of $\widehat{\theta}_N$, and Efron (1982, p. 13) states that it is a better estimator of the variance of $\widehat{\theta}_N$ than of the variance of $\widehat{\theta}_J$, it can also be used as an estimator of the variance of $\widehat{\theta}_J$. The standard error of $\widehat{\theta}_J$ is then estimated by $\sqrt{\widehat{\sigma}_J^2}$.

If $m > 1$, the sample can be divided into g mutually exclusive groups of size m, where $g = N/m$. Of course this is only possible when g is an integer. Now, call the estimator based on the total sample from which group j is removed $\widehat{\theta}_{(j)}$ and the according bias estimator (2.34) $\widehat{\text{bias}}_{(j)}$. The average of the estimators $\widehat{\theta}_{(j)}$ $(j = 1, \ldots, g)$ is called $\widehat{\theta}_{(.)}$ and the grouped-jackknife estimator of the bias ($\widehat{\text{bias}}_J$) is the average of the estimators $\widehat{\text{bias}}_{(j)}$. The grouped-jackknife bias-corrected estimator of θ is $\widehat{\theta}_N - \widehat{\text{bias}}_J$, which is equal to

$$\widehat{\theta}_J = g\widehat{\theta}_N - (g-1)\widehat{\theta}_{(.)}, \tag{2.39}$$

which is completely analogous to (2.37). The according grouped-jackknife variance estimator is also completely similar to the ungrouped-jackknife case (2.38). It is given by

$$\widehat{\sigma}_J^2 = \frac{g-1}{g} \sum_{j=1}^{g} \left(\widehat{\theta}_{(j)} - \widehat{\theta}_{(.)} \right)^2 . \tag{2.40}$$

It is also possible to have g groups of possibly different sizes. In this case, let m_j be the size of group j and let $\widehat{\theta}_{(j)}$ be the estimator reestimated from the sample from which group j was removed, and let $\widehat{\text{bias}}_{(j)}$ be the accompanying estimator of the bias of $\widehat{\theta}_N$ from (2.34). An unweighted bias estimator is now

$$\widehat{\text{bias}}_J = \frac{1}{g} \sum_{j=1}^{g} \widehat{\text{bias}}_{(j)}.$$

The according bias-corrected estimator of θ is

$$\begin{aligned} \widehat{\theta}_J &= \widehat{\theta}_N - \widehat{\text{bias}}_J \\ &= g\widehat{\theta}_N \left(\frac{1}{g} \sum_{j=1}^{g} \frac{N/m_j}{g} \right) - (g-1) \left[\frac{1}{g} \sum_{j=1}^{g} \left(\frac{\frac{N}{m_j} - 1}{g-1} \right) \widehat{\theta}_{(j)} \right], \end{aligned} \tag{2.41}$$

which reduces to the standard grouped-jackknife bias-corrected estimator if the group sizes are all equal. The unweighted estimator of the variance of $\widehat{\theta}_J$ is

$$\widehat{\sigma}_J^2 = \frac{g-1}{g} \sum_{j=1}^{g} \left(\widehat{\theta}_{(j)} - \widehat{\theta}_{(.)} \right)^2 . \tag{2.42}$$

The formulas for the grouped jackknife estimators in the case that the group sizes are unequal are experimental. The bias-corrected estimator (2.41) should be relatively unbiased, though possibly not optimally efficient. It is unclear whether the variance estimator (2.42) is approximately correct. More research is needed to shed light on these issues.

2.8.2 The bootstrap

The *bootstrap* was introduced by Efron (1979) as an alternative to the jackknife.

The idea of the bootstrap is that the empirical distribution function is a consistent estimator of the distribution function in the population. Let Z be a random variable with distribution function F, and let $\{z_1, z_2, \ldots, z_N\}$ be a random sample of size N from F. Now, the empirical distribution function $\widehat{F}_N$ in some point z is the proportion of z_i that are smaller than or equal to z:

$$\widehat{F}_N(z) = \frac{\#\{i : 1 \leq i \leq N | z_i \leq z\}}{N}. \tag{2.43}$$

If Z has a multivariate distribution, this formula has an obvious generalization and all subsequent formulas will also have obvious generalizations. It is known (e. g., Mood et al., 1974, p. 507) that, as $N \longrightarrow \infty$, $\widehat{F}_N(z) \longrightarrow F(z)$.

Let θ be a parameter associated with the distribution F, $\theta = \theta(F)$, and let $\widehat{\theta}$ be an estimator of θ from a sample, $\widehat{\theta} = \theta(z_1, z_2, \ldots, z_N) = \theta(\widehat{F}_N)$. The idea of the bootstrap is now to simulate the sampling and estimation process, where samples are drawn from $\widehat{F}_N$, which is completely known once the original sample is obtained. In the simulation, the distribution $\widehat{F}_N$ plays the role of F and $\widehat{\theta}$ plays the role of θ: Simulation samples $\{z_1^*, z_2^*, \ldots, z_N^*\}$ are drawn from $\widehat{F}_N$ and $\widehat{\theta}$ is estimated by θ^* in the same way θ was estimated by $\widehat{\theta}$.

Because $\widehat{F}_N \longrightarrow F$, it is assumed that the properties of the estimator θ^* based on the distribution $\widehat{F}_N$ give information about the properties of $\widehat{\theta}$ based on the distribution F. For example, the bias of θ^* based on the distribution $\widehat{F}_N$ is taken as an estimator of the bias of $\widehat{\theta}$ based on the distribution F. It has been proved by many authors that this approach works in many cases, that is, that it leads to consistent estimators of the properties of $\widehat{\theta}$ (e. g., Putter, 1994). The actual implementation of the bootstrap is quite simple: Drawing samples from $\widehat{F}_N$ is equivalent to drawing samples with replacement from $\{z_1, z_2, \ldots, z_N\}$.

The bootstrap is now implemented as follows: B bootstrap samples $\{z_{b1}^*, z_{b2}^*, \ldots, z_{bN}^*\}$, $b = 1, \ldots, B$, are drawn from $\widehat{F}_N$, that is, these samples are drawn with replacement from $\{z_1, z_2, \ldots, z_N\}$. From each of the B samples, the parameter $\widehat{\theta}$ is estimated, thereby obtaining B estimators θ_b^*, $b = 1, \ldots, B$. Now the expectation of θ^* (given $\widehat{F}_N$) is estimated by the mean of the estimators θ_b^*, namely, $\theta_{(.)}^* = \sum_{b=1}^B \theta_b^*/B$. The variance of θ^* (given $\widehat{F}_N$) is estimated by the variance of the estimators θ_b^*, namely, $\widehat{\mathrm{var}}(\theta^*) = \sum_{b=1}^B (\theta_b^* - \theta_{(.)}^*)^2/B$.

The bias of $\widehat{\theta}$ is estimated by the (estimated) bias of θ^*:

$$\widehat{\mathrm{bias}}_B = \widehat{\mathrm{bias}}(\theta^*) = \theta_{(.)}^* - \widehat{\theta}, \tag{2.44}$$

and the bias-corrected estimator of θ is therefore

$$\begin{aligned} \widehat{\theta}_B &= \widehat{\theta} - \widehat{\mathrm{bias}}_B \\ &= 2\widehat{\theta} - \theta_{(.)}^*. \end{aligned} \tag{2.45}$$

The variance of $\widehat{\theta}$ is simply estimated by the variance of θ_b^*:

$$\widehat{\mathrm{var}}_B = \widehat{\mathrm{var}}(\theta^*) = \frac{1}{B} \sum_{b=1}^{B} \left(\theta_b^* - \theta_{(.)}^* \right)^2. \tag{2.46}$$

The parametric bootstrap

The bootstrap as described above can also be termed the *nonparametric bootstrap*, because the distribution the bootstrap samples are drawn from is the nonparametric empirical distribution function $\widehat{F}_N$. Frequently, however, it is assumed that F is a specific distribution $F(\phi)$, only depending on a parameter (or parameter vector) ϕ, which may or may not be the same parameter as θ. Then, if ϕ is estimated by $\widehat{\phi}$, F can also be estimated by $\widetilde{F}_N = F(\widehat{\phi})$, instead of $\widehat{F}_N$. If the distributional assumption about F is correct, this *parametric* empirical distribution function will generally be a better (more efficient) estimator of F.

The *parametric bootstrap* is defined exactly analogous to the nonparametric bootstrap, except that bootstrap samples are drawn from $\widetilde{F}_N$ instead of $\widehat{F}_N$. This means that no longer samples are drawn with replacement from the original data, but from a generally more smooth distribution function. Hence, the values of the z^*_{bi} in the bootstrap sample will usually not be values also encountered in the original sample.

For example, if it is assumed that F is a normal distribution function with mean μ and variance σ^2, then bootstrap samples are drawn from a normal distribution with mean $\overline{x}$ and variance s^2, where $\overline{x}$ and s^2 are the mean and variance of the original sample.

2.8.3 Resampling regression models

Consider a simple linear regression model

$$y = \alpha + \beta x + \varepsilon,$$

where ε is a normally distributed error term with mean zero and variance σ^2. Suppose that a sample $\{(y_1, x_1), \ldots, (y_N, x_N)\}$ is available. Then parameter estimates $\widehat{\alpha}$, $\widehat{\beta}$, and $\widehat{\sigma}^2$ can be obtained. Now, if x is considered a random variable, nonparametric bootstrap samples can be easily obtained by resampling complete *cases*: Bootstrap samples $\{(y^*_1, x^*_1), \ldots, (y^*_N, x^*_N)\}$ consist of pairs (y^*_i, x^*_i) that are also elements of the original sample, that is, for each $i = 1, \ldots, N$, there exists a j, $1 \leq j \leq N$, such that $(y^*_i, x^*_i) = (y_j, x_j)$. Then, the parameters can be estimated from each bootstrap sample and bias-corrected estimates can be obtained, as well as an estimate of the covariance matrix of the estimator, using the formulas from section 2.8.2.

The implementation of the parametric bootstrap depends on whether a specific distribution of x is assumed. If x is regarded as a random variable with an *unspecified* distribution, the parametric bootstrap should start with drawing *nonparametric* bootstrap samples of x. If, on the other hand, a *specific* distribution of x is assumed, for example, a normal distribution with mean μ and variance σ^2_x, then the parametric bootstrap starts with drawing *parametric* bootstrap samples of x, for example, samples from a normal distribution with mean $\overline{x}$ and variance s^2_x, which are the estimates of μ and σ^2_x from the original sample.

Given a bootstrap sample $\{x^*_1, \ldots, x^*_N\}$ of x, the parametric bootstrap draws a sample $\{\varepsilon^*_1, \ldots, \varepsilon^*_N\}$ of ε from a normal distribution with mean zero and variance $\widehat{\sigma}^2$, where $\widehat{\sigma}^2$ is the estimate of σ^2 from the original sample. Then, a bootstrap sample $\{y^*_1, \ldots, y^*_N\}$ of y is computed from the following equation:

$$y^*_i = \widehat{\alpha} + \widehat{\beta} x^*_i + \varepsilon^*_i, \tag{2.47}$$

where $\widehat{\alpha}$ and $\widehat{\beta}$ are the estimates of α and β from the original sample.

The situation is different if x is regarded as a fixed (design) variable, chosen by the experimentor. This happens, for example, if x is the dose of some drug administered to

rats by the experimentor. Then each bootstrap sample should have exactly the same x values, that is, $x_i^* = x_i$ for each i in each bootstrap sample. The *parametric* bootstrap is in this case simply obtained by (2.47), with $x_i^* = x_i$. The *nonparametric* bootstrap is in this case, however, completely different from the nonparametric bootstrap with random x. In this case, first, the errors are estimated from the original sample by

$$\widehat{\varepsilon}_i = y_i - \widehat{\alpha} - \widehat{\beta} x_i. \tag{2.48}$$

Then, bootstrap samples $\{\varepsilon_1^*, \ldots, \varepsilon_N^*\}$ are drawn from $\{\widehat{\varepsilon}_1, \ldots, \widehat{\varepsilon}_N\}$, and bootstrap samples of y are obtained analogously to (2.47):

$$y_i^* = \widehat{\alpha} + \widehat{\beta} x_i + \varepsilon_i^*. \tag{2.49}$$

Then, bootstrap estimates of the parameters and bootstrap estimates of the covariance matrix of the parameters are obtained in the usual way (e. g., Efron, 1982, pp. 35–36).

The *jackknife* can also be implemented straightforwardly in regression models: One complete case is removed from the sample for each $\widehat{\theta}_{(i)}$ for the ungrouped jackknife, or a group of complete cases is removed for each $\widehat{\theta}_{(j)}$ for the grouped jackknife. The jackknife bias-corrected estimators and the jackknife estimators of the covariance matrix of the parameters are obtained straightforwardly (e. g., Efron, 1982, pp. 18–19).

The bootstrap and jackknife methods discussed here for regression models are the standard implementations as, for example, discussed by Efron (1982). These have some drawbacks, and therefore, alternative resampling methods have been proposed that have some advantages, for example, that they are robust to heteroskedasticity. A thorough discussion can be found in Wu (1986).

2.8.4 Resampling multilevel models

Because multilevel analysis is based on regression analysis, resampling methods for multilevel models can be based on resampling methods for regression models. The methods of section 2.8.3 can, however, not straightforwardly be applied to multilevel models, because the usual jackknife and bootstrap theory requires that the different observations be independently distributed. This is not the case with multilevel analysis, where the observations within the same Level-2 unit are dependent.

Another difference between regression analysis and multilevel analysis is that in multilevel analysis, there can be variables measured at all levels. In the two-level case, for example, there are variables describing the Level-1 units and (possibly) variables describing the Level-2 units. This implies that resampling can be performed at two levels.

Consider two-level data. A straightforward implementation of the (ungrouped) *jackknife* would be to eliminate one observation from one Level-2 unit at the time to obtain a jackknife sample. This resampling scheme is exactly equivalent to the resampling scheme of the standard ungrouped jackknife of section 2.8.1. Another possibility is to implement the grouped jackknife. With the grouped jackknife, it is most logical to use the Level-2 units as groups. The Level-2 units may have different sizes, and therefore, the grouped jackknife with unequal group sizes (2.41)–(2.42) should be used. The theoretical properties of these estimators are currently not known. Moreover, the grouped and ungrouped jackknife are based on the assumption that the observations are independent, which is not the case in multilevel analysis. Furthermore, the jackknife estimators for regression analysis may also be nonoptimal (Wu, 1986). Therefore, the jackknife estimators in multilevel analysis are experimental and may not be consistent. Further research will be needed to

obtain information about the properties of these estimators. For that purpose, they are implemented as options in MLA.

The *parametric bootstrap* can be easily implemented in multilevel analysis. If the X_j and W_j variables are considered fixed in (2.1) and (2.2), bootstrap samples $\{y_{b1}^*, \ldots, y_{bJ}^*\}$ can be obtained in the following way. First, for each $j = 1, \ldots, J$, draw a bootstrap Level-2 error vector u_j^* from a normal distribution with mean zero and covariance matrix $\widehat{\Theta}$. Then, draw a bootstrap Level-1 error vector ε_j^* from a normal distribution with mean zero and covariance matrix $\widehat{\sigma}_\varepsilon^2 I_{N_j}$. Finally, the bootstrap sample of y is obtained from

$$\beta_j^* = W_j \widehat{\gamma} + u_j^* \tag{2.50}$$

and

$$y_j^* = X_j \beta_j^* + \varepsilon_j^*. \tag{2.51}$$

Then, bias-corrected bootstrap estimators and bootstrap estimators of the covariance matrix of the parameters are obtained in the usual way. This is the parametric bootstrap that is implemented in MLA. Note that this simulation option is also provided by ML3 (Prosser et al., 1991). It is also possible to derive a parametric bootstrap estimator in case the X and W variables are considered random. This is analogous to (2.47), but it is not implemented in MLA.

For the *nonparametric bootstrap*, several situations can be studied. If the X and W variables can be considered *fixed*, then, analogously to regression analysis, the *errors* have to be estimated. As explained in section 2.5, the shrunken residuals (2.26) and (2.28) can be used as estimators of the Level-2 and Level-1 errors, respectively. A drawback of these errors may be that their variances are less than the variances in the population. When, however, sample sizes at both levels increase, this difference diminishes. But, alternatively, the *raw residuals* (2.24)–(2.25) can be used instead of the shrunken residuals.

Unlike in regression analysis, the estimated residuals in multilevel analysis do not necessarily have a zero mean. Therefore, the means are subtracted first. Otherwise, the possibly nonzero mean of the errors would necessarily lead to biased estimators of the constant. Once (centered) estimates $\{\widehat{u}_j\}$, $j = 1, \ldots, J$, and $\{\widehat{\varepsilon}_{ij}\}$, $j = 1, \ldots, J$, $i = 1, \ldots, N_j$, of the errors are obtained, nonparametric bootstrap samples $\{u_j^*\}$, $j = 1, \ldots, J$, and $\{\varepsilon_{ij}^*\}$, $j = 1, \ldots, J$, $i = 1, \ldots, N_j$ are drawn, and nonparametric bootstrap samples of y are obtained from (2.50) and (2.51). Then, estimators can be obtained in the usual way, and bootstrap bias-corrected estimators and standard errors can be obtained straightforwardly. This bootstrap procedure of resampling from estimated errors is called the *error bootstrap*.

Whether the shrunken or raw residuals are to be preferred in bootstrapping multilevel models is unclear yet. They are both implemented as options in MLA. It is also unclear whether these bootstrap methods are satisfactory, or other bootstrap methods should be used instead (Wu, 1986).

If the X and W variables are considered *random*, *nonparametric* bootstrap samples can be drawn by resampling complete *cases*. This is, however, somewhat more complicated than in regression analysis, because the hierarchical structure of the data should be respected. The bootstrap samples can be drawn in the following way. First, a sample of size J is drawn with replacement from the *Level-2 units*. This gives a sample j_k^*, $k = 1, \ldots, J$ of Level-2 unit numbers and accompanying Level-2 variables $W_{j_k^*}$. Then for each k, a nonparametric bootstrap sample of complete cases from the (original) unit

$j = j_k^*$ is drawn, giving $\{(y_{ik}^*, X_{ik}^*), k = 1, \ldots, J, i = 1, \ldots, N_{j_k^*}\}$. This is called the *cases bootstrap* for both levels.

It is also possible to draw bootstrap samples from the Level-2 units only, keeping all the y's, X's, and W's fixed once a Level-2 unit is drawn. This is useful when the data within the unit can not be considered a simple random sample, for example, with repeated measures data or families. Then, a complete Level-2 unit is (temporarily) regarded as a single observation and bootstrap samples are drawn from these observations. With repeated measures, this implies that for each subject that is drawn in the bootstrap sample, the data for all the timepoints are exactly the same as in the original sample. For a family, this means that the complete family is kept together, and that, once the family is drawn in the bootstrap sample, mother, father, and children are all part of the bootstrap sample, and, for example, the mother can not be drawn twice within the same Level-2 unit.

On the other hand, it is also possible to keep the Level-2 units fixed, and draw bootstrap samples only from the Level-1 units within each Level-2 unit. This can be useful when the Level-2 units can not be considered a simple random sample, for example, when several (prespecified) countries are compared and people within each country are drawn randomly. Then, in the bootstrap samples, all countries are present once, just as in the original sample. Bootstrap samples are drawn from complete cases within each country.

Once bootstrap samples are drawn, bootstrap bias-corrected estimators and bootstrap standard errors can be obtained straightforwardly.

The three possible methods for drawing bootstrap samples from complete cases discussed above are implemented as options in `MLA` as well. It will depend on the nature of the data which one is most fit for a particular application.

2.9 Missing data

Missing data are a frequently occurring phenomenon. For instance, in repeated measures designs, the points in time at which the different subjects are measured may not be the same, or the number of points in time the subjects are measured may differ. This situation leads to missing time-points, that is, all time-specific variables of a subject are missing at some point in time. However, the time-invariant variables (such as sex) are, of course, known. This situation is easily handled by a multilevel model, in which the subjects are the Level-2 units, and the time-points are the Level-1 units. As was discussed for a usual multilevel model, the number of Level-1 units may be different for different Level-2 units, and so the missing timepoints give no problems. An example of repeated measures is given in chapter 4.

If, however, in a multilevel model, be it an application in repeated measures or not, for some Level-1 unit, some Level-1 variables are measured, but others are not (or for some Level-2 unit, some Level-2 variables are measured, but others are not), there are missing values that can not be handled by the standard model. If only output variables are missing, the `EM` algorithm provides a standard way of dealing with the missing values in a satisfactory way. If, however, some exogenous (X and/or Z) variables are missing, the `EM` algorithm can not be used straightforwardly, because it requires that the joint distribution of the exogenous and the endogenous (output) variables is known. Standard multilevel modeling only assumes that the conditional distribution of the output variables given the exogenous variables is known. This poses severe complications.

If the amount of data that is missing is relatively small, standard ad-hoc solutions to the missing-data problem can be used, such as *listwise deletion* (deletion of cases with one

or more missing values), *pairwise deletion* (computation of "sufficient" statistics, such as covariances, on the basis of all available information for the variables in question), *mean substitution* (substitution of the mean of the observed values of a variable for a missing value on that variable), or other substitution methods. All these methods have their advantages and drawbacks and none is fully satisfactory, especially when the number of missing values is large.

In the current version of `MLA`, no specific means of missing-data handling are implemented. Listwise deletion and several forms of substitution can be done by the user before the data set is processed by `MLA`. Pairwise deletion can not be done, because the program requires raw data. In principle, pairwise deletion could be done within the program, but this is not implemented (yet).

Chapter 3

Input

In this chapter the input of the MLA program is explained. A simple introduction to multilevel models is given in Chapter 1. A discussion of estimation and other relevant theory concerning the multilevel model implemented in MLA can be found in Chapter 2.

MLA Version 1.0b runs as a stand-alone batch program. It uses an input file and an output file as parameters. This means that, in DOS, the program can be started by the command

MLA *input-file-name output-file-name*

where *input-file-name* is the name of the file that contains the input and *output-file-name* is the name of the file in which the output of the program will be saved. Both files are simple text files (ascii). The output file will be explained in the next chapter. The input file will be considered here.

The input file consists of statements, which are case insensitive. Every statement begins with a slash and a keyword (e. g., /TITLE). Every keyword may be abbreviated, but it must be at least of length three to be recognized (e. g., /TIT). Other text following the keyword and/or leading spaces will be ignored. The rest of the statements must follow on lines below the keyword and should precede the next statement. These lines are called substatements and may also consist of one or more keywords (e. g., file). The last statement to be read is the /END statement. All other statements, and corresponding substatements, may appear in any order (but before the /END statement if they are to be executed). Finally, comments, preceded by a percent sign (%), may appear throughout the input file. All text on a line, after and including the percent sign, will serve as comment and is ignored as program input.

In the following, all statements and substatements implemented are discussed and illustrated with small examples. In Chapter 4, where we focus on the program output, complete examples are provided.

3.1 /TITLE (optional)

Following the keyword /TITLE, the first non-blank line contains the title for the analysis. Although the statement is optional, it is highly recommended. Moments after the analysis all may seem clear, but after a few months you may have no idea what you have done. The title may be your only clue. You may also enrich your input file with comments. In contrast to comments, the title is repeated on top of every part of the output.

Example:

```
/TITLE
  MLA example 1: analysis of variance
```

3.2 /DATA (required)

The `/DATA` statement contains information about the data file. This statement has four substatements, three of which are required. The `file` substatement gives the name of the data file, `variables` the number of variables in the data file, `id1` the (optional) variable number of the Level-1 identifier variable, and `id2` the variable number of the Level-2 identifier variable.

Example:

```
/DATA
  file = sesame.dat  % data set from Glasnapp and Poggio (1985)
  vars = 3           % total of three variables
  id2 = 1            % Level-2 identification given by first variable
```

3.2.1 file (required)

This substatement indicates the name of the data file. The name is given after the equals sign and must satisfy the usual `DOS` conventions on filenames. If the file is in the current directory the complete pathname is not necessary. The file itself is a free-field formatted numbers-only `ascii` file. This means that values of variables must be separated by at least one blank. A case may consist of more than one line. Cases must be sorted by the Level-2 identifier variable (see below).

3.2.2 variables (required)

The `variables` substatement specifies the number of variables in the data file. Because the data file is a free-field formatted file and one case may consist of more than one line, this is necessary information for the program to determine when to start a new case.

3.2.3 id1 (optional)

With this substatement, a case number variable can be given. This can be useful in those situations where the output gives specific information about cases at the first level. The substatement is otherwise equal to the `id2` substatement (see below). If omitted, the order in which the Level-1 units are read from the data file is used as identification.

3.2.4 id2 (required)

One of the variables in the data file must contain a code (number) that identifies the Level-2 units. This may be a group number or, in case of repeated measurements, a subject number. The number is essential for a correct discrimination of the Level-2 units. Level-1 units are interchangeable within a Level-2 unit. A Level-1 identifier variable is not necessary. The variable number has to follow the keyword `id2` and it must indicate the position of the identifier variable in the data file. The variable number must be at least 1 and less than or equal to the number of variables, indicated in the `variables` substatement.

3.3 /MODEL (required)

The /MODEL statement is followed by a set of equations that specify the model that has to be estimated. Every equation must be on a single line. There is only one Level-1 equation, but there may be one or more Level-2 equations. The order in which the Level-1 and Level-2 equations appear is arbitrary. The terms used in the Level-1 equation are:

- Vi = variable i, which is the i-th variable in the data file. Vi may be either indicating the outcome variable or a predictor variable.
- Bi = beta component i (β_{ij}, the ith element of a typical β_j, cf. equation 2.1). At Level-1 these are the regression coefficients that seem to be outcome variables at Level-2 (cf. equation 2.2).
- E = the Level-1 random term (ε). This term is considered to be a residual or error term. The variance of this term has to be estimated from the data.

The Level-2 equations partly consist of the same terms, but also of specific Level-2 equation terms:

- Bi = beta component i, corresponding with the Level-1 regression coefficient. At this level, however, Bi can be viewed as an outcome variable.
- Gi = gamma component i (γ_i). These are the fixed parameters to be estimated in the multilevel model.
- Vi = one of the variables from the data file (as explained above). In this case, it is a Level-2 predictor variable. It means that this variable is considered to have the same value for all Level-1 units within a particular Level-2 unit. To be certain that this is the case, for each Level-2 variable the average is computed over all Level-1 units within that particular Level-2 unit.[1]
- Ui = Level-2 random term i (u_{ij}, the ith element of a typical u_j). As with the first level, this component is considered a residual or error term, but now for the second level. The second level may have more than one error term: one for each Level-2 equation (i. e., for each β element). The variances and the covariances of these terms have to be estimated from the data.

Example:

```
/MODEL
  B1 = G1 + G2*V6 + U1  % random intercepts, dependent on level-2 predictor
  B2 = G3 + G4*V6 + U2  % random slopes, dependent on the same level-2 predictor
  V4 = B1 + B2*V5 + E   % level-1 equation, dependent on level-1 predictor
```

In the equations each term is followed by a number (except for the Level-1 random term E). For the Vi term this number is the variable number, the position of the variable in the data file (e. g., V4, the fourth variable in the data file). The other terms only use a number for identification, without any additional meaning (e. g., G3, one of the fixed parameters). The Bi terms have meaning in the equations of both levels. Every equation consists of one term before and at least one term after the equals sign. The minimal specification of a model is:

[1] Note that this feature may be used to create an aggregated Level-1 variable, serving as a Level-2 predictor variable, simply by specifying a Level-1 variable as a Level-2 variable as well.

```
/MODEL
  B1 = G1      % fixed intercept
  V4 = B1 + E  % level-1 variation
```

or

```
/MODEL
  B1 = U1      % random intercept
  V4 = B1 + E  % level-1 variation
```

As shown above, terms on the right hand side of the equations are connected by plus signs. A variable and a corresponding parameter are connected by an asterisk (`*`). This is used to connect a fixed parameter and an observed predictor variable in Level-2 equations and to connect a Level-1 regression coefficient and an observed predictor variable in the Level-1 equation. In Chapter 4, several variations of the two-level model will be presented and discussed in more detail.

3.4 /CONSTRAINTS (optional)

`MLA` has a limited option for imposing parameter constraints. Parameters to be estimated may be constrained to a certain value. Constraints are imposed as: "parameter = value". This feature is only implemented for the `FIML` estimation part. It is simply ignored for the various `OLS` estimators.

Example:

```
/CONSTRAINTS
  G1 = 1.0  % fix component G1 to 1.0
  U1 = 0.5  % fix level-2 variance of U1 to 0.5
```

Values must be specified as floating-point numbers. Covariances are specified by connecting the appropriate Level-2 residual terms by an asterisk.

Example:

```
/CONSTRAINTS
  U1*U2 = 0.0  % fix level-2 covariance U1*U2 to 0.0
```

3.5 /SIMULATION (optional)

Several options for simulation are available in `MLA`. These include the jackknife and three versions of the bootstrap (Efron, 1982). Theoretical details concerning the implementation of these resampling methods for the two-level model can be found in Chapter 2.

With the substatements provided with the `/SIMULATION` statement, one can choose between the different kinds of simulation (using the keyword `kind`), and specify special simulation features (using the keywords `method`, `type` and `resample`). Additional features are the number of replications and the initial seed for the random number generator (`replications` and `seed`). Finally, one can specify a separate output file for intermediate results of the simulation (`file`).

Example:

```
\SIMULATION
  kin = bootstrap  % use simulation method bootstrap
  met = error      % resample from error vectors
  typ = raw        % use raw residuals as error vectors
  res = 1          % only resample level-1 units
  rep = 200        % repeat simulation 200 times
  see = 1041245    % start with random seed 1041245
  fil = boot.out   % write simulation results to boot.out
```

3.5.1 `kind` (required)

With this substatement the user can choose from two options, namely `bootstrap` and `jackknife` simulation.

Both types of simulation work as follows.

- obtain a (new) sample
- repeat the analysis
- save the (new) estimates

These three steps, together called a replication, are repeated a number of times. Afterwards, bias-corrected estimates of model parameters and nonparametric estimates of standard errors are computed. These estimates are computed from the set of saved (bootstrap or jackknife) estimates and the original maximum likelihood estimates.

The *bootstrap*, introduced by Efron (1979), differs from the *jackknife*, the nonparametric technique proposed by Quenouille (1949), in the way a new sample is obtained. The choice between bootstrap or jackknife resampling also determines the way the final simulation estimates are computed. More details can be found in Chapter 2.

3.5.2 `method`

This substatement specifies the method of bootstrap to be performed. It is required whenever `kind = bootstrap`. One can choose between three different methods: `error`, `cases`, and `parametric`. The three methods differ in the way the bootstrap sample is obtained.

`error`

This method resamples the elements of the Level-1 and Level-2 error vectors. Subsequently a new outcome or dependent variable is computed using these error vectors, the original predictor or independent variables and their corresponding `FIML` parameter estimates. The way in which the Level-1 and Level-2 error terms are estimated from the "total" `FIML` residuals is discussed in Chapter 2.

`cases`

Using this method a bootstrap sample is created by resampling the original data. Thus, complete cases are randomly drawn (with replacement) from the original cases. The procedure follows the nested structure in the data, by a nested resampling of cases: Level-2 units are randomly drawn (with replacement) and cases within a particular drawn unit are resampled. It is also possible to resample only complete Level-2 units, where the Level-1 units within a sampled Level-2 units are the same as in the original data set (which is useful for repeated measures data), or to resample only Level-1 units within Level-2 units, where the Level-2 units are the same as in the original sample, but the Level-1 units within each Level-2 units are resampled (useful when there are few Level-2 units and many Level-1 units in each Level-2 unit, such in studies with many subjects from a few countries).

`parametric`

This method computes a new outcome or dependent variable using the original predictor variables, their corresponding `FIML` parameter estimates and a set of random Level-1 and Level-2 error terms. These terms are obtained as follows: New Level-1 errors are drawn from a normal distribution with mean zero and variance $\hat{\sigma}^2$, which is the `FIML` estimate of the Level-1 variance component. New Level-2 errors are drawn from a (multivariate) normal distribution with zero mean vector and covariance matrix $\widehat{\Theta}$, which contains the `FIML` estimates of the Level-2 variance components.

3.5.3 `type`

The substatement `type` is only required whenever the substatement `kind = bootstrap` is used in combination with `method = error`. The `type` substatement specifies the type of estimation that is used to determine the Level-1 and Level-2 residuals. One can choose between `raw` and `shrunken`. More details can be found in Chapter 2.

3.5.4 `resample` (optional)

The substatement `resample` offers the user the choice at which level units will be resampled. The default is `0`, which means that at both levels units will be resampled. If `kind = jackknife`, or `kind = bootstrap` and `method = cases`, the user may choose `1` or `2`, which means that only Level-1 units or only Level-2 units will be resampled, respectively. The kind of nested structure in the data will determine which choice is appropriate. For instance, with repeated measures (Level-1) nested within individuals (Level-2), it is probably not useful to resample Level-1 units with the cases bootstrap.

3.5.5 `replications` (optional)

Using the substatement `replications` the number of bootstrap replications is specified. It must be an integer value between 1 and 32767 ($2^{15} - 1$). The default value is 300 and this number is usually considered sufficient, although Markus (1994) suggests 1000 in another context.

3.5.6 `seed` (optional)

For diagnostic purposes, one can provide an initial number (seed) for the random number generator. This is specified by the substatement `seed`. Using the same initial seed, the simulation samples will be identical. The seed value must be an integer between 1 and 1,073,735,823 ($2^{31} - 1$). If results from bootstrap analyses are to be reported it is advised to save the seeds.

3.5.7 `file` (optional)

Results of the simulation analysis can be written to a file. Using the substatement `file`, a filename may be specified. Filenames must satisfy the ususal `DOS` conventions on filenames. For each replication, the following results are written to the file (in `ascii`, space separated):

1. global information
    - replication number

- seed
- number of iterations until convergence
- the minimum of the -2 log likelihood function

2. estimation results: pairs containing

   - estimate
   - standard error

   of each parameter. The parameters are in the following order: σ_ε^2, γ_1, ..., γ_p, Θ_{11}, Θ_{21}, Θ_{22}, Θ_{31}, ..., Θ_{qq}, where p is the dimension of γ and q is the dimension of each β_j.

The estimation results are thus repeated "`replications`" times and displayed with a maximum of eight values per line (four estimates and their corresponding standard errors). The results of the simulation analysis are used to compute the final bootstrap and jackknife estimates. The results of a replication are not taken into account when the algorithm did not converge or when the estimate or its standard error was fixed to zero because it reached the edge of its parameter space. Further elaboration concerning this subject can be found both in the previous and in the next chapter.

3.6 /TECHNICAL (optional)

The `/TECHNICAL` statement provides useful possibilities to alter the estimation process. It concerns the estimation method (`method`), minimization stop criteria, like the maximum number of iterations (`maximum number of iterations`) and two convergence criteria (`fconvergence` and `pconvergence`), the critical p-value for the display of outliers, and the possibility of writing intermediate iteration results to disk (`file`). If this statement and subsequent substatements are not specified, the program will run using default values.

Example:

```
/TECHNICAL
  met = fiml      % estimation method fiml
  max = 10        % maximum number of iterations equals 10
  fco = 0.00001   % function convergence set to 0.00001
  pco = 0.0001    % parameter convergence set to 0.0001
  out = 0.01      % critical p-value for outlier display
  fil = tech.out  % technical results will be written to tech.out
```

3.6.1 method (optional)

The substatement `method` provides the opportunity to set the estimation method. One can choose between `FIML` and `REML`[2]. `FIML` is the default method and represents full information maximum likelihood estimation. `REML` is restricted maximum likelihood estimation. Both procedures are described in Chapter 2 and in Appendix A.

[2] not implemented yet

3.6.2 `maximum number of iterations` (optional)

The default value of `max` is 20. This number should be sufficient for reaching convergence if the sample size is large enough and/or the number of parameters to be estimated is not too large. Changing the convergence criteria (see below) can make it necessary to raise the maximum number of iterations. The value must be an integer between 1 and 32767 ($2^{15} - 1$).

3.6.3 `fconvergence` (optional)

The substatement `fconvergence` refers to function convergence. After each iteration the new function value is compared to the previous function value. The obtained difference is compared to a `fconvergence` related value. If

$$\frac{|F_{i-1} - F_i|}{(|F_i| + |F_{i-1}|)/2} \leq \texttt{fconvergence},$$

convergence is said to have been reached. In this formula, F_i is the function value after the ith iteration. The first part of the formula represents the ratio between the difference of two successive function values and the mean of these values. The default value of `fconvergence` is 0.001 and permitted values range from 1.0 to 1.0E-16.

3.6.4 `pconvergence` (optional)

The substatement `pconvergence` refers to parameter convergence. After each iteration the parameter vector is compared with its predecessor by computing a vector of differences. Using ν_i as the norm of this vector after the ith iteration, convergence is said to have been reached when

$$\nu_i \leq \texttt{pconvergence}.$$

The default value of `pconvergence` is 0.001 and permitted values range from 1.0 to 1.0E-16. The use of this substatement has no influence on the estimation process while simulating, because the loss of speed resulting from its use.

3.6.5 `outliers` (optional)

With this substatement, the critical p-value for outlier display can be set. The keyword is `outliers` and an example is given below.

```
/TECHNICAL
  outliers = 0.25
```

On the extremes, `outliers = 0.0` will show no outliers at all, while `outliers = 1.0` will show all available cases, being both Level-1 and Level-2 units.

3.6.6 `file` (optional)

The technical output can be written to a separate file. The file is specified after the `file` substatement under the `/TECHNICAL` statement and must satisfy the usual `DOS` conventions on filenames. The file will contain the iteration numbers and the parameter estimates (in the same order as in section 3.5.7) after each iteration. Depending on the number of parameters, multiple lines of parameter estimates will be displayed.

3.7 /OUTPUT (optional)

The /OUTPUT statement gives the user control over the output. Not all output is optional. The default output consists of a title page, an echo of the input, the maximum likelihood estimates (FIML), and system information. Output for the simulation analysis is generated whenever the /SIMULATION statement is used. Additional output is controlled by keywords following the /OUTPUT statement. The keywords must be separated by spaces or commas and may take up more than one line. The keywords will be shortly explained below. A more profound elaboration follows in the chapter on output. Most theory underlying the different parts of the output can be found in Chapter 2.

Example:

```
\OUTPUT
  input,              % display digested input statements
  des,                % display variable descriptive statistics
  out,ols,res,pos,dia % display all other output
```

input

The input information is digested and displayed in two parts. A required and an optional part. Here, a single equation is displayed (similar to (2.3)) and input can be checked. After the input information, a short table of contents of the output is displayed. It explains which part of the output gives which information.

descriptives

All data variables are used to obtain simple summary statistics. The sample sizes of the different levels are displayed followed by two blocks of information. The first block displays mean, stddev, variance, skewness, kurtosis, and K-S Z (the latter with its significance level denoted by ???), respectively. Computational formulas are given in Chapter 2. The second block contains seven percentiles of the variables. These are the 0th (minimum), 5th (P5), 25th (Q1), 50th (median), 75th (Q3), 95th (P95), and 100th (maximum) percentiles, respectively.

outcomes

The Level-2 outcomes consist of ordinary least squares estimates per Level-2 unit. Estimates of the regression coefficients and estimates of the error variance, including their standard errors, t-ratios and exceedance probabilities of the t-ratios per Level-2 unit are displayed in separate blocks with their Level-2 unit number and Level-2 unit size.

olsquares

This part contains the ordinary least squares estimates for the fixed (Gi) and random (variances and covariances of Ui and E) parameters. A regression analysis is performed, ignoring grouping. For the error variance two estimates are displayed, the one-step (E(1)) and two-step (E(2)) estimates, corresponding to (2.14) and (2.22), respectively.

technical

The technical information output consists of four columns. The iteration number is in the first column. In the second column the -2 log likelihood is displayed. The third column

shows the norm of the difference of the parameter vector at the current iteration and the parameter vector at the previous iteration. That is, the differences between the current and previous values of the parameters are squared and summed, and the square root of the resulting summation is reported. The last column shows the norm of the gradient vector. When the `technical` keyword is not used, only the final information is displayed as part of the maximum likelihood information part.

`residuals`

For the first level, three different types of residuals are displayed, namely the total, raw, and shrunken residuals. The Level-2 residuals are the raw and shrunken residuals for every random Level-2 component. Computational formulas are given in Chapter 2.

`posterior`

Displayed are the posterior means (2.29) based on the full information maximum likelihood estimates.

`diagnostics`

For diagnostic purposes outliers are reported. There are two kinds of outliers, one for each level. See chapter 2 for details.

Chapter 4

Output

The output of MLA consists of a single text file, the second parameter in the statement that starts program execution. The file is divided into parts. One part may take more than one page. Before each new part, a pagebreak is inserted.

This chapter will elaborate on the MLA output file We will illustrate the output using several example analyses. These will stretch from a simple analysis of variance to a bootstrap analysis for a complicated two-level model. It is not our intention to give extensive examples of case studies. The examples will give insight in how to use MLA for different analyses and glance at specific parts of the output.

4.1 Analysis of variance

To illustrate how to run an ANOVA using MLA, we consider part of the Sesame Street data set. The original set from Glasnapp and Poggio (1985) is used in Stevens (1990) for an analysis of covariance. In the first example with this data set, we only use two variables from the set, which originally included 12 background variables and 8 achievement variables for 240 subjects. The first 3 sites of the original 5 sites are used on both pretest and posttest. Only the achievement variable measuring knowledge of numbers is considered here. The series was viewed in between the pretest and posttest. The series was meant to teach pre-school skills to 3 to 5 year old children.

An analysis of variance is performed on these data with MLA. Level-2 (j) indicates the site and Level-1 (i) the children. The model to be estimated is

$$Y_{ij} = \gamma + u_j + \varepsilon_{ij}, \tag{4.1}$$

where γ is the overall mean on the posttest score, u_j is the Level-2 deviation from γ, or Level-2 error component, and ε_{ij} is the Level-1 deviation from $\gamma + u_j$, the average score of unit j, also called the Level-1 error component. Equation (4.1) can be divided into two separate equations, one for each level:

$$\begin{aligned} Y_{ij} &= \beta_j + \varepsilon_{ij}, \\ \beta_j &= \gamma + u_j. \end{aligned}$$

In this way, the deviations or error components for the different levels are easily seen. These equations are also the equations that are to be used in MLA to specify the model. Along with the other statements, the input file is as follows:

```
/TITLE
```

```
  MLA example 1: analysis of variance
/DATA
  file = sesame.dat
  vars = 3
  id2 = 1
/MODEL
  b1 = g1 + u1
  v3 = b1 + e
/OUTPUT
  inpu,desc,olsq
/END
```

All output contains the MLA title page. It is the first part of the output. It only supplies information about the name and origin of the program. It is not possible to leave this part out.

```
            MMMM        MMMMM  LLLL  AAAAAAAA
           MMMMM      MMMMMM  LLLL  AAAAAAAAAA
          MMMM M    MMMMMMM  LLLL  AAAA    AAAA
         MMMM  MM MMM MMMM  LLLL  AAAA      AAAA
        MMMM   MMMM  MMMM  LLLL  AAAA        AAAA
       MMMM    MM   MMMM  LLLL  AAAAAAAAAAAAAAAAAA
      MMMM     M   MMMM  LLLL  AAAAAAAAAAAAAAAAAAAA
     MMMM         MMMM  LLLL  AAAA              AAAA
    MMMM         MMMM  LLLL  AAAA                AAAA
   MMMM         MMMM  LLLL                       AAAA
  MMMM         MMMM  LLLLLLLLLLLLLLLLLLLLLLLLL  AAAA
 MMMM         MMMM  LLLLLLLLLLLLLLLLLLLLLLLLLL  AAAA
                                                AAAA
 MULTILEVEL ANALYSIS FOR TWO LEVEL DATA          AAAA
                                                  AAAA
 VERSION 1.0b                                      AAAA
                                                    AAAA
 DEVELOPED BY                                        AAAA
   FRANK BUSING                                       AAAA
   ERIK MEIJER                                         AAAA
   RIEN VAN DER LEEDEN                                  AAAA
                                                         AAAA
 PUBLISHED BY                                             AAAA
   LEIDEN UNIVERSITY                                       AAAA
   FACULTY OF SOCIAL AND BEHAVIOURAL SCIENCES               AAAA
   DEPARTMENT OF PSYCHOMETRICS AND RESEARCH METHODOLOGY      AAAA
   WASSENAARSEWEG 52                                          AAAA
   P.O. BOX 9555                                               AAAA
   2300 RB LEIDEN                                               AAAA
   THE NETHERLANDS                                               AAAA
   PHONE +31 (0)71-273761                                         AAAA
   FAX   +31 (0)71-273619                                          AAAA
```

Except for the title page and the optional `input` part, every part contains a header. The header is always the same and is made of two lines of standard text and the title of the analysis, supplied by the user. For this first example, it reads:

```
 MLA (U)   MULTILEVEL ANALYSIS FOR TWO LEVEL DATA   VERSION 1.0b  09-10-1994
 COPYRIGHT 1993-1994 LEIDEN UNIVERSITY   ALL RIGHTS RESERVED        PART  2

 MLA EXAMPLE 1: ANALYSIS OF VARIANCE
```

The second part of the output contains an echo of the input-file statements. This part is always included in an output file.

```
 INPUTFILE STATEMENTS

  1  /TITLE
  2    MLA example 1: analysis of variance
  3  /DATA
```

```
 4    file = sesame.dat
 5    vars = 3
 6    id2 = 1
 7  /MODEL
 8    b1 = g1 + u1
 9    v3 = b1 + e
10  /OUTPUT
11    inpu,desc,olsq
12  /END

  12 LINES WERE READ FROM INPUTFILE "EXAMPLE1.IN"
```

The third part is the first optional part of the output. It is triggered by the `input` keyword under the `/OUTPUT` statement. It contains extra information about the input and the output. Specifically, the input statements are digested and re-displayed and a short table of contents of the output is given.

```
INPUT INFORMATION

  REQUIRED

    NAME OF DATAFILE      : SESAME.DAT
    NUMBER OF VARIABLES   : 3
    LEVEL-2 ID. COLUMN    : 1
    MODEL SPECIFICATION   : B1=G1+U1
                            V3=B1+E
    SINGLE EQUATION       : V3=E+G1+U1

  OPTIONAL

    TITLE OF ANALYSIS     : MLA EXAMPLE 1: ANALYSIS OF VARIANCE
    ESTIMATION METHOD     : FULL INFORMATION MAXIMUM LIKELIHOOD

OUTPUT INFORMATION

  PART  CONTENTS

     1  TITLE PAGE
     2  INPUTFILE STATEMENTS
     3  INPUT INFORMATION
     4  DATA DESCRIPTIVES
     5  ORDINARY LEAST SQUARES ESTIMATES
     6  FULL INFORMATION MAXIMUM LIKELIHOOD ESTIMATES
     7  SYSTEM INFORMATION
```

The single equation shows the integration of the Level-2 equations and the Level-1 equation in the same way as in Chapter 2, Equations (2.1), (2.2) and (2.4). It is displayed directly below the model specification. The output information displays the different parts in the output. Default as well as optional output parts are mentioned in two columns, one for the part number and one for the contents.

The fourth part consists of the data descriptives, optionally given by the use of the keyword `descriptives` under the `/OUTPUT` statement.

These statistics are displayed in two major blocks, and are preceded by the number of Level-1 and Level-2 units.

```
DATA DESCRIPTIVES

# LEVEL-1 UNITS = 179
# LEVEL-2 UNITS = 3

VAR      MEAN    STDDEV  VARIANCE  SKEWNESS  KURTOSIS     K-S Z   PROB(Z)

  1      2.02      0.83      0.70     -0.04     -1.57      3.18      0.00
  2     21.37     10.92    119.25      0.72     -0.16      1.45      0.03
```

```
  3     31.02      12.89     166.19      -0.07      -1.13       1.37       0.05

VAR    MINIMUM         P5         Q1     MEDIAN         Q3        P95    MAXIMUM

  1       1.00       1.00       1.00       2.00       3.00       3.00       3.00
  2       4.00       7.00      13.50      19.00      28.00      44.00      52.00
  3       0.00      10.00      20.00      31.00      42.50      51.00      54.00
```

The first variable is the Level-2 identifier variable. The second and third variables are the score on the pretest and the posttest, respectivily. Formulas can be found in Section 2.2.

Part 5 gives `OLS` estimates. (This part is also optional. The user must supply the keyword `olsquares` in the `/OUTPUT` statement.) As described in Chapter 2, ordinary least squares estimation yields two different estimates for the Level-1 variance component, σ^2, one by ignoring the hierarchical data structure and one using this structure. These are both displayed in Part 5 of the output. The one-step estimate is labeled `E(1)` and the two-step estimate is labeled `E(2)`. `U1*U1` gives the variance estimate for the Level-2 variance component `U1`.

```
ORDINARY LEAST SQUARES ESTIMATES

FIXED PARAMETERS

   LABEL       ESTIMATE            SE

      G1      31.016760       0.963540

RANDOM PARAMETERS

   LABEL       ESTIMATE            SE

    E(1)     166.185111      17.615587

   U1*U1      29.469076      24.061400

    E(2)     136.503030      14.469292


E(1) : ONE-STEP ESTIMATE OF SIGMA SQUARED (IGNORING GROUPING)
E(2) : TWO-STEP ESTIMATE OF SIGMA SQUARED
SEE DOCUMENTATION FOR FURTHER ELABORATION ON THESE SUBJECTS
```

As can be seen, the overall mean (`G1`) equals the mean of Variable 3, the score on the posttest (31.02). Ignoring grouping will result in 166.19 for σ^2. Using the two-step procedure lowers the estimate to 136.50 and also gives an estimate of the variance of u_j (29.47).

Part 6 contains the `FIML` estimates. This part is default and appears in all output. Compared to the previous ordinary least squares estimates part, `T`-values and probabilities for `T` are given. Here, unlike for the `OLS` estimates, these are theoretically justified.

```
FULL INFORMATION MAXIMUM LIKELIHOOD ESTIMATES

FIXED PARAMETERS

   LABEL       ESTIMATE            SE             T        PROB(T)

      G1      31.322433       3.123586         10.03         0.0000

RANDOM PARAMETERS

   LABEL       ESTIMATE            SE             T        PROB(T)
```

```
   U1*U1      26.935304      23.900162          1.13         0.2597

       E     138.833164      14.799662          9.38         0.0000

INTRA-CLASS CORRELATION = 26.9353 / ( 138.8332 + 26.9353 ) = 0.1625


CONVERGENCE CRITERION REACHED

# ITERATIONS = 5
-2*LOG(L)    = 1398.626571
```

Whenever there are residuals associated with the grand mean, the intra-class correlation is computed and given just below the `FIML` estimates. The formula for the intra-class correlation is

$$\rho = \frac{\sigma^2}{\tau + \sigma^2} \tag{4.2}$$

and in `MLA` notation,

$$\rho = \frac{\mathtt{E}}{\mathtt{U1} * \mathtt{U1} + \mathtt{E}} \tag{4.3}$$

If the `technical` keyword is omitted from the `/OUTPUT` statement a short description of the final iteration results is given in the `FIML` part. Here, convergence is reached in 5 iterations and yields a `-2*LOG(L)` value of 1398.63.

The final part of the output contains some system information. The format of the date is `DD-MM-YYYY` and for the time `HH:MM:SS`. The elapsed time is in $\frac{1}{100}$ seconds (`HH:MM:SS:HH`). The program is terminated correctly in about a quarter of a second, as can be seen in the seventh and final, default, part of the output.

```
SYSTEM INFORMATION

              START      FINISH        ELAPSED

  DATE    19-12-1994  19-12-1994
  TIME      12:52:47    12:52:47    00:00:00:16


PROGRAM TERMINATED CORRECTLY
```

4.2 Analysis of covariance

For the next example the same Sesame Street data set is used. Now, an analysis of covariance is performed on these data with `MLA`. The model to be estimated is

$$Y_{ij} = \gamma_1 + \gamma_2 X_{ij} + u_j + \varepsilon_{ij}, \tag{4.4}$$

where γ_1 is the overall mean, X_{ij} is the covariate, u_j is the Level-2 error component and ε_{ij} is the Level-1 error component. Equation (4.4) can be divided into separate equations, one equation for Level-1 and in this case two Level-2 equations:

$$\begin{aligned} Y_{ij} &= \beta_{1j} + \beta_{2j} X_{ij} + \varepsilon_{ij}, \\ \beta_{1j} &= \gamma_1 + u_j, \\ \beta_{2j} &= \gamma_2. \end{aligned}$$

Along with the other statements, the input file is as follows:

```
/TITLE
  MLA example 2: analysis of covariance
/DATA
  file = sesame.dat
  vars = 3
  id2 = 1
/MODEL
  b1 = g1 + u1
  b2 = g2
  v3 = b1 + b2*v2 + e
/OUTPUT
  olsq
/END
```

Compared to the previous example, a fixed parameter (G2) is added in the OLS-estimates part. This is the regression coefficient of the Level-1 covariate containing the pretest score.

```
 ORDINARY LEAST SQUARES ESTIMATES

 FIXED PARAMETERS

    LABEL       ESTIMATE             SE

       G1      14.672451       1.621040
       G2       0.764871       0.067590

 RANDOM PARAMETERS

    LABEL       ESTIMATE             SE

     E(1)      96.968087      10.307591

    U1*U1       7.217027       5.892678

     E(2)      88.980063       9.458474


 E(1) : ONE-STEP ESTIMATE OF SIGMA SQUARED (IGNORING GROUPING)
 E(2) : TWO-STEP ESTIMATE OF SIGMA SQUARED
 SEE DOCUMENTATION FOR FURTHER ELABORATION ON THESE SUBJECTS
```

The parameter estimate for the regression coefficient of the covariate is also added to the FIML output part. The additional T-value and PROB(T) indicate that the pretest variable explains a significant part of the variance of the posttest variable ($T = 10.18$, $PROB(T) = 0.0000$).

```
 FULL INFORMATION MAXIMUM LIKELIHOOD ESTIMATES

 FIXED PARAMETERS

    LABEL       ESTIMATE             SE              T        PROB(T)

       G1      16.196937       2.226470           7.27         0.0000
       G2       0.699891       0.068761          10.18         0.0000

 RANDOM PARAMETERS

    LABEL       ESTIMATE             SE              T        PROB(T)

    U1*U1       6.766701       6.759616           1.00         0.3168

        E      89.831188       9.576026           9.38         0.0000

 INTRA-CLASS CORRELATION = 6.7667 / ( 89.8312 + 6.7667 ) = 0.0701
```

```
CONVERGENCE CRITERION REACHED

# ITERATIONS = 7
-2*LOG(L)    = 1318.217264
```

Entering the covariate into the analysis is justified, because it has a statistically significant non-zero effect. The same justification could be made with the use of the likelihood-ratio test. This test is based on the fact that the difference between minus two times the loglikelihood function value (`-2*LOG(L)`) of two nested models follows a chi-square distribution with the number of degrees of freedom equal to the difference in the number of free parameters. The two models (Example 1 and Example 2) are nested and the likelihood-ratio test can be applied. The difference between the function values is approximately $1399 - 1318 = 81$, and the degrees of freedom is equal to 1. The likelihood-ratio test indicates that the effect is highly significant.

4.3 Repeated measures analysis

The rat data set used for the repeated measures example has been analyzed by a number of investigators. The first use of these data with multilevel analysis appeared in (Strenio, Weisberg, & Bryk, 1983). The rat data consist of the weights of ten rats. These rats were measured five times with four week intervals from birth. Also included in the model is the weight of each rat's mother (`V2`). Divided into two levels, the equations are given by

$$
\begin{aligned}
Y_{ij} &= \beta_{1j} + \beta_{2j} X_{ij} + \varepsilon_{ij}, \\
\beta_{1j} &= \gamma_1 + \gamma_2 * W_j + u_{1j}, \\
\beta_{2j} &= \gamma_3 + \gamma_4 * W_j + u_{2j},
\end{aligned}
$$

where X_{ij} (`V5`) is the age (in weeks, divided by 4, minus 2, so that it is in deviation of the mean) of the rat. W_j (`V2`) represents the weight of the mother. The input file for the repeated measures example is as follows.

```
/TITLE
  MLA example 3; repeated measures analysis
/DATA
  file = rat.dat
  vars = 4
  id2 = 3
/MODEL
  b1 = g1 + g2*v2 + u1
  b2 = g3 + g4*v2 + u2
  v1 = b1 + b2*v4 + e
/OUTPUT
  outc,post
/END
```

For each rat a multiple regression analysis is performed and displayed in the Level-2 outcomes part. This part is optional and displayed through the use of the `outcomes` keyword in the `/OUTPUT` statement.

```
LEVEL-2 OUTCOMES: ORDINARY LEAST SQUARES ESTIMATES PER LEVEL-2 UNIT

UNIT    SIZE          B1         SE(B1)           T        PROB(T)

   1       5    111.4000         4.9044       22.71         0.0000
   2       5    120.2000         2.9967       40.11         0.0000
   3       5    119.8000         6.7621       17.72         0.0000
   4       5    103.4000         3.8018       27.20         0.0000
```

```
    5       5       100.0000         3.2701         30.58         0.0000
    6       5        99.0000         4.4505         22.24         0.0000
    7       5        93.0000         5.5281         16.82         0.0000
    8       5       113.6000         1.6391         69.31         0.0000
    9       5        90.4000         4.5284         19.96         0.0000
   10       5       121.0000         2.4549         49.29         0.0000

MEAN               107.1800

UNIT    SIZE            B2         SE(B2)             T        PROB(T)

    1       5        28.8000         3.4679          8.30         0.0000
    2       5        28.1000         2.1190         13.26         0.0000
    3       5        36.3000         4.7816          7.59         0.0000
    4       5        27.2000         2.6882         10.12         0.0000
    5       5        23.4000         2.3123         10.12         0.0000
    6       5        29.3000         3.1470          9.31         0.0000
    7       5        25.6000         3.9090          6.55         0.0000
    8       5        19.7000         1.1590         17.00         0.0000
    9       5        23.6000         3.2021          7.37         0.0000
   10       5        25.6000         1.7359         14.75         0.0000

MEAN                26.7600

UNIT    SIZE         SIGMA2     SE(SIGMA2)             T        PROB(T)

    1       5       120.2667        98.1973          1.22         0.2207
    2       5        44.9000        36.6607          1.22         0.2207
    3       5       228.6333       186.6783          1.22         0.2207
    4       5        72.2667        59.0055          1.22         0.2207
    5       5        53.4667        43.6554          1.22         0.2207
    6       5        99.0333        80.8604          1.22         0.2207
    7       5       152.8000       124.7607          1.22         0.2207
    8       5        13.4333        10.9683          1.22         0.2207
    9       5       102.5333        83.7181          1.22         0.2207
   10       5        30.1333        24.6038          1.22         0.2207

MEAN                91.7467
```

In the next part we can see that both `G2` and `G4` indicate that the mother's weight has a positive effect on the rat's weight. The rat's weight starts higher and rises faster with a heavier mother.

```
FULL INFORMATION MAXIMUM LIKELIHOOD ESTIMATES

FIXED PARAMETERS

   LABEL       ESTIMATE            SE             T        PROB(T)

      G1      18.873660      18.784897          1.00         0.3150
      G2       0.545101       0.115348          4.73         0.0000
      G3       2.967709      10.597305          0.28         0.7794
      G4       0.146866       0.065072          2.26         0.0240

RANDOM PARAMETERS

   LABEL       ESTIMATE            SE             T        PROB(T)

   U1*U1      18.585604      17.183843          1.08         0.2794
   U2*U1      -6.598562       6.911578         -0.95         0.3397
   U2*U2       2.580007       5.765943          0.45         0.6545

       E      91.746665      23.688887          3.87         0.0001

INTRA-CLASS CORRELATION = 18.5856 / ( 91.7467 + 18.5856 ) = 0.1685


CONVERGENCE CRITERION REACHED
```

```
# ITERATIONS = 10
-2*LOG(L)    = 376.262296
```

The posterior means may be compared with the Level-2 outcomes. As can be seen, the posterior means tend to be shrunken towards the grand mean, and therefore have less variance than the Level-2 outcomes.

```
POSTERIOR MEANS

UNIT            B1

   1       111.2096
   2       123.6838
   3       118.5373
   4       103.2717
   5       102.5805
   6        98.5713
   7        92.8730
   8       109.4713
   9        96.1403
  10       115.4610

MEAN       107.1800

UNIT            B2

   1        28.0714
   2        31.7001
   3        31.3653
   4        26.1328
   5        25.9444
   6        25.8216
   7        23.7186
   8        23.1345
   9        26.1874
  10        25.5239

MEAN        26.7600
```

4.4 Multilevel analysis

In 1988, the National Center for Education Statistics of the U.S. Department of Education collected data on amount of homework done and scores on math tests from students of more than 1000 schools. The subset from this National Education Longitudinal Study (NELS) of 1988 data used for this example consists of ten manually selected schools, containing 260 students from Public (coded 1) and Private (coded 0) schools. These data were also used as an example analysis in Kreft and Van Der Leeden (1994). The model equations are given by

$$
\begin{aligned}
Y_{ij} &= \beta_{1j} + \beta_{2j} X_{ij} + \varepsilon_{ij}, \\
\beta_{1j} &= \gamma_1 + \gamma_2 * W_j + u_{1j}, \\
\beta_{2j} &= \gamma_3 + \gamma_4 * W_j + u_{2j},
\end{aligned}
$$

where Y_{ij} represents the score on the math test (`V9`), X_{ij} (`V5`) is the amount of homework done, and W_j is the variable indicating the type of school (Public or Private) (`V17`). The following input file shows the application of the `/TECHNICAL` statement as well, where the maximum number of iterations is raised to 100 and both convergence criteria are set to 0.00001.

```
/TITLE
  MLA example 4: multilevel analysis
/DATA
  file = nels.dat
  vars = 17
  id2 = 1
/MODEL
  b1 = g1 + g2*v17 + u1
  b2 = g3 + g4*v17 + u2
  v9 = b1 + b2*v5 + e
/TECHNICAL
  maxiter = 100
  fconv = 0.00001
  pconv = 0.00001
/OUTPUT
  tech
/END
```

The special output given contains the intermediate iteration results. These are the results during the full maximum likelihood estimation. Starting with the values from the two-step `OLS` estimation, we can see that the first value is already rather good. More about the difference between first iteration and final estimates in two level models can be found in Van Der Leeden and Busing (1994). The estimation process stopped after 14 iterations. The difference between the last two `-2*LOG(L)` values was less than the user provided `fconvergence` (0.00001).

```
TECHNICAL ITERATION INFORMATION

ITER     -2*LOG(L)     NORM(dP)      NORM(G)

   1  1749.8685307    1.9340571   11.8087864
   2  1749.6327546    0.6335839    1.1931032
   3  1749.6156005    0.0572362    0.7851259
   4  1749.6110404    0.1167412    0.7776501
   5  1749.4998336    0.4502308    0.7366149
   6  1749.4862306    0.0849216    0.7000794
   7  1749.4491604    0.1256636    0.6790682
   8  1749.4462601    0.1138294    0.0593836
   9  1749.4441677    0.1133145    0.0583135
  10  1749.4439035    0.0301898    0.0290881
  11  1749.4439029    0.0014868    0.0020701
  12  1749.4439027    0.0046646    0.0002624
  13  1749.4439026    0.0005529    0.0002900
  14  1749.4439026    0.0000003    0.0000056


CONVERGENCE CRITERION REACHED

NORM(dP) : LENGTH OF DIFFERENCE BETWEEN SUCCESSIVE PARAMETER-VECTORS
NORM(G)  : LENGTH OF GRADIENT-VECTOR
SEE DOCUMENTATION FOR FURTHER ELABORATION ON THESE SUBJECTS
```

The following part gives the `FIML` estimates.

```
FULL INFORMATION MAXIMUM LIKELIHOOD ESTIMATES

FIXED PARAMETERS

   LABEL       ESTIMATE            SE             T        PROB(T)

      G1      59.098244       6.547975          9.03         0.0000
      G2     -15.827270       6.925261         -2.29         0.0223
      G3       1.108726       4.648499          0.24         0.8115
      G4       0.922201       4.916968          0.19         0.8512

RANDOM PARAMETERS
```

```
    LABEL       ESTIMATE            SE             T        PROB(T)

    U1*U1      39.862888      20.314503          1.96         0.0497
    U2*U1     -28.697577      14.153063         -2.03         0.0426
    U2*U2      21.390296      10.257549          2.09         0.0370

        E      42.782546       3.902927         10.96         0.0000

  INTRA-CLASS CORRELATION = 39.8629 / ( 42.7825 + 39.8629 ) = 0.4823
```

As can be concluded from the output, the interaction term `G4` is not significant. A model without this term might be preferred because it is a more parsimonious model. It does, however, not alter the significant negative effect of the school type.

4.5 Simulation study

In the Junior School Project (Inner London Education Authority, 1987), the following variables were collected: Mathematics Achievement in Years 1 through 3, an ability measure (score on the Ravens test in Year 1), and sex. There are 48 classes present from 36 different schools with a total of 887 children. The input file shown below represents a bootstrap study with resampling from the shrunken residuals. Resampling from both levels is used with 200 replications. Together with the other input, the input file is as follows.

```
/TITLE
  MLA example 5: simulation study
/DATA
  file = jsp.dat
  vars = 7
  id2 = 1
/MODEL
  b1 = g1 + u1
  b2 = g2
  b3 = g3
  v5 = b1 + b2*v4 + b3*v3 + e
/SIMULATION
  kind = bootstrap
  method = cases
  resample = 1
  replications = 200
  seed = 1
/END
```

The `FIML` estimates are given below. In fact, this model is an analysis of covariance with two covariates at the first level (`G2` and `G3`). Thus only one random estimate for the second level is specified (`U1`).

```
  FULL INFORMATION MAXIMUM LIKELIHOOD ESTIMATES

  FIXED PARAMETERS

    LABEL       ESTIMATE            SE             T        PROB(T)

       G1      15.251835       0.896721         17.01         0.0000
       G2       0.592560       0.032978         17.97         0.0000
       G3       1.272573       0.443152          2.87         0.0041

  RANDOM PARAMETERS

    LABEL       ESTIMATE            SE             T        PROB(T)
```

```
   U1*U1       4.049940       1.184081          3.42         0.0006

       E      27.852020       1.359013         20.49         0.0000

INTRA-CLASS CORRELATION = 4.0499 / ( 27.8520 + 4.0499 ) = 0.1269


CONVERGENCE CRITERION REACHED

# ITERATIONS = 9
-2*LOG(L)    = 5527.578950
```

After 200 bootstrap replications, there are no replications that are incorrect (i. e., with inadmissible parameter values or non-convergence). The final bootstrap estimates that were computed are given below.

```
BOOTSTRAP ESTIMATES

  # REPLICATIONS         = 200
  # CORRECT REPLICATIONS = 200

FIXED PARAMETERS

   LABEL       ESTIMATE            SE

      G1      15.254733       0.044857
      G2       0.592006       0.007469
      G3       1.271495       0.028049

RANDOM PARAMETERS

   LABEL       ESTIMATE            SE

   U1*U1       4.013912       0.300021

       E      27.954714       0.385577
```

This bootstrap simulation took about half a minute on a 486-DX2 66MHz. This amounts to 6 replications per second. This means that there do not have to be any obstacles for using a high number of replications. Literature suggests many different numbers, ranging from 100 replications to 1000 replications, taking 17 seconds to 3 minutes computer time.

Appendix A

Technical Appendix

In this appendix, the theory of maximum likelihood estimation used in the `MLA` program will be discussed in detail. Other authors, such as Bryk and Raudenbush (1992) and (Longford, 1987) give some technical detail as well, but much is left to the reader. We think, however, that it is useful to explain in much more detail what is actually done in the program, and this appendix serves this purpose. In this appendix, the (minus-log-)likelihood function and its gradient function are derived, as well as computationally more efficient formulas of them. The asymptotic covariance matrix of the maximum likelihood estimators and computationally efficient formulas for it are derived and the explicit imposition of implicit constraints in the model is discussed.

A.1 The model and the likelihood function

To find maximum likelihood estimates, we start with the model (2.4):

$$y_j = X_j\gamma + Z_j u_j + \varepsilon_j \tag{A.1}$$

$$\varepsilon_j \sim N(0, \sigma^2 I_{N_j}) \tag{A.2}$$

$$u_j \sim N(0, \Theta), \tag{A.3}$$

where y_j is a vector with the endogenous variable for the N_j Level-1 units in Level-2 unit j, X_j is an $N_j \times p$ matrix of exogenous variables for the Level-1 units in Level-2 unit j, and Z_j is an $N_j \times q$ matrix of exogenous variables for the Level-1 units in Level-2 unit j. The p-vector γ is a vector of fixed regression coefficients, the q-vector u_j is a vector of random regression coefficients in Level-2 unit j, and the N_j-vector ε_j is a vector of residuals of the Level-1 units in Level-2 unit j. It is assumed that ε_j and u_j are independent of each other and independent of $\varepsilon_{j'}$ and $u_{j'}$, where $j' \neq j$.

From the model equations (A.1)–(A.3), it is found that, conditional on X_j and Z_j, y_j is normally distributed and the expectation and covariance matrix of y_j are

$$Ey_j = X_j\gamma \tag{A.4}$$

$$\begin{aligned} V_j &= E(y_j - X_j\gamma)(y_j - X_j\gamma)' \\ &= \sigma^2 I_{N_j} + Z_j \Theta Z_j'. \end{aligned} \tag{A.5}$$

Consequently, the probability density of y_j is

$$f(y_j) = \frac{1}{(2\pi)^{N_j/2}(\det V_j)^{1/2}} e^{-\frac{1}{2}(y_j - X_j\gamma)' V_j^{-1}(y_j - X_j\gamma)},$$

so that the contribution of Level-2 unit j to the minus-log-likelihood function is

$$\begin{aligned} L_j &= -\log f(y_j) \\ &= \frac{N_j}{2}\log(2\pi) + \frac{1}{2}\log\det V_j + \frac{1}{2}(y_j - X_j\gamma)'V_j^{-1}(y_j - X_j\gamma) \end{aligned}$$

and the minus-log-likelihood function for the whole sample is

$$\begin{aligned} L &= \sum_{j=1}^{J} L_j \\ &= \frac{N}{2}\log(2\pi) + \frac{1}{2}\sum_{j=1}^{J}\log\det V_j + \frac{1}{2}\sum_{j=1}^{J}(y_j - X_j\gamma)'V_j^{-1}(y_j - X_j\gamma), \end{aligned} \tag{A.6}$$

where J is the number of Level-2 units, and N is the total number of Level-1 units, $N = \sum_{j=1}^{J} N_j$, where N_j is the number of Level-1 units in Level-2 unit j. This is the function that has to be minimized with respect to the parameters to obtain maximum likelihood estimators. To minimize this function, the program uses the *Broyden-Fletcher-Goldfarb-Shanno* (BFGS) minimization method (see, e.g., Press et al., 1986), which uses the gradient of the function to be minimized.

In section A.3, computationally efficient formulas for the function and the gradient will be derived. In section A.4, the asymptotic covariance matrix of the estimators will be derived. In section A.5, a reparametrization of the model will be discussed, in which the restriction of positive (semi-)definiteness of covariance matrices is explicitly imposed. But first, in the next section, some useful notation, matrices, and formulas will be introduced.

A.2 Some useful formulas

First, we define the matrix

$$G_j = I_q + Z_j'Z_j\Theta/\sigma^2. \tag{A.7}$$

This matrix will be used frequently in the following.

The inverse of V_j. Maddala (1977, p. 446), states the following formula:

$$(A + BDB')^{-1} = A^{-1} - A^{-1}B(D^{-1} + B'A^{-1}B)^{-1}B'A^{-1},$$

where A and D are square nonsingular matrices and B is a matrix of appropriate dimensions. This formula can also be written as

$$\begin{aligned} (A + BDB')^{-1} &= A^{-1} - A^{-1}B(D^{-1} + B'A^{-1}B)^{-1}B'A^{-1} \\ &= A^{-1} - A^{-1}B[(I + B'A^{-1}BD)D^{-1}]^{-1}B'A^{-1} \\ &= A^{-1} - A^{-1}BD(I + B'A^{-1}BD)^{-1}B'A^{-1}. \end{aligned} \tag{A.8}$$

By defining $A = \sigma^2 I_{N_j}$, $B = Z_j$, and $D = \Theta$, it follows that V_j can be written as $A + BDB'$. Consequently, the inverse of V_j can be found from equation (A.8):

$$\begin{aligned} V_j^{-1} &= \sigma^{-2}I_{N_j} - \left(\sigma^{-2}I_{N_j}\right)Z_j\Theta\left[I_q + Z_j'\left(\sigma^{-2}I_{N_j}\right)Z_j\Theta\right]^{-1}Z_j'\left(\sigma^{-2}I_{N_j}\right) \\ &= \sigma^{-2}I_{N_j} - \sigma^{-4}Z_j\Theta\left(I_q + Z_j'Z_j\Theta/\sigma^2\right)^{-1}Z_j' \\ &= \sigma^{-2}I_{N_j} - \sigma^{-4}Z_j\Theta G_j^{-1}Z_j'. \end{aligned} \tag{A.9}$$

The determinant of V_j. Based on Maddala (1977, pp. 446–447), the following formula for the determinant of a partitioned matrix can be derived:

$$\begin{aligned}\det\begin{pmatrix} A & B \\ C & D \end{pmatrix} &= \det\begin{pmatrix} A & B \\ C & D \end{pmatrix}\begin{pmatrix} I & -A^{-1}B \\ 0 & I \end{pmatrix} \\ &= \det\begin{pmatrix} A & 0 \\ C & D - CA^{-1}B \end{pmatrix} \\ &= \det A \ \det(D - CA^{-1}B),\end{aligned}$$

where A and D are square nonsingular matrices, and B and C are matrices of appropriate orders.

Similarly,

$$\begin{aligned}\det\begin{pmatrix} A & B \\ C & D \end{pmatrix} &= \det\begin{pmatrix} A & B \\ C & D \end{pmatrix}\begin{pmatrix} I & 0 \\ -D^{-1}C & I \end{pmatrix} \\ &= \det D \ \det(A - BD^{-1}C).\end{aligned}$$

Consequently,

$$\det A \ \det(D - CA^{-1}B) = \det D \ \det(A - BD^{-1}C). \tag{A.10}$$

Now, define $A = I_q$, $B = Z_j'$, $C = -Z_j\Theta$, and $D = \sigma^2 I_{N_j}$. The matrix V_j can now be written as $V_j = D - CA^{-1}B$, and the determinant of A is 1. Consequently, using equation (A.10),

$$\begin{aligned}\det V_j &= \det A \det(D - CA^{-1}B) \\ &= \det D \det(A - BD^{-1}C) \\ &= \det(\sigma^2 I_{N_j}) \det[I_q - (Z_j')(\sigma^2 I_{N_j})^{-1}(-Z_j\Theta)] \\ &= (\sigma^2)^{N_j} \det(I_q + Z_j'Z_j\Theta/\sigma^2) \\ &= (\sigma^2)^{N_j} \det G_j. \end{aligned} \tag{A.11}$$

The factor $Z_j'V_j^{-1}$. In the following, the factor $Z_j'V_j^{-1}$ will frequently pop up. This factor can be written in a computationally more efficient form:

$$\begin{aligned}Z_j'V_j^{-1} &= Z_j'(\sigma^{-2}I_{N_j} - \sigma^{-4}Z_j\Theta G_j^{-1}Z_j') \qquad \text{(from (A.9))} \\ &= \sigma^{-2}Z_j' - \sigma^{-4}Z_j'Z_j\Theta G_j^{-1}Z_j' \\ &= \sigma^{-2}Z_j' - \sigma^{-2}(Z_j'Z_j\Theta/\sigma^2)G_j^{-1}Z_j' \\ &= \sigma^{-2}Z_j' - \sigma^{-2}(I_q + Z_j'Z_j\Theta/\sigma^2 - I_q)G_j^{-1}Z_j' \\ &= \sigma^{-2}Z_j' - \sigma^{-2}(G_j - I_q)G_j^{-1}Z_j' \qquad \text{(from (A.7))} \\ &= \sigma^{-2}Z_j' - \sigma^{-2}Z_j' + \sigma^{-2}G_j^{-1}Z_j' \\ &= \sigma^{-2}G_j^{-1}Z_j'. \end{aligned} \tag{A.12}$$

From (A.12) it follows that

$$Z_j'V_j^{-1}Z_j = \sigma^{-2}G_j^{-1}Z_j'Z_j \tag{A.13}$$

and

$$\begin{aligned}Z_j'V_j^{-2}Z_j &= \sigma^{-2}G_j^{-1}Z_j'V_j^{-1}Z_j \\ &= \sigma^{-4}G_j^{-2}Z_j'Z_j. \end{aligned} \tag{A.14}$$

The traces of V_j^{-1} and V_j^{-2}. From equation (A.9), we find

$$\begin{aligned}
\mathrm{tr}V_j^{-1} &= \mathrm{tr}\left(\sigma^{-2}I_{N_j} - \sigma^{-4}Z_j\Theta G_j^{-1}Z_j'\right) \\
&= \sigma^{-2}N_j - \sigma^{-4}\mathrm{tr}(Z_j\Theta G_j^{-1}Z_j') \\
&= \sigma^{-2}N_j - \sigma^{-4}\mathrm{tr}(Z_j'Z_j\Theta G_j^{-1}) \\
&= \sigma^{-2}N_j - \sigma^{-2}\mathrm{tr}[(Z_j'Z_j\Theta/\sigma^2)G_j^{-1}] \\
&= \sigma^{-2}N_j - \sigma^{-2}\mathrm{tr}[(I_q + Z_j'Z_j\Theta/\sigma^2 - I_q)G_j^{-1}] \\
&= \sigma^{-2}N_j - \sigma^{-2}\mathrm{tr}[(G_j - I_q)G_j^{-1}] \\
&= \sigma^{-2}N_j - \sigma^{-2}\mathrm{tr}(I_q - G_j^{-1}) \\
&= \sigma^{-2}N_j - \sigma^{-2}q + \sigma^{-2}\mathrm{tr}G_j^{-1}. \qquad \text{(A.15)}
\end{aligned}$$

Similarly, using (A.9), (A.15), and (A.12),

$$\begin{aligned}
\mathrm{tr}V_j^{-2} &= \mathrm{tr}\left[\left(\sigma^{-2}I_{N_j} - \sigma^{-4}Z_j\Theta G_j^{-1}Z_j'\right)V_j^{-1}\right] \\
&= \sigma^{-2}\mathrm{tr}V_j^{-1} - \sigma^{-4}\mathrm{tr}(Z_j\Theta G_j^{-1}Z_j'V_j^{-1}) \\
&= \sigma^{-4}(N_j - q) + \sigma^{-4}\mathrm{tr}G_j^{-1} - \sigma^{-4}\mathrm{tr}\left[Z_j\Theta G_j^{-1}(\sigma^{-2}G_j^{-1}Z_j')\right] \\
&= \sigma^{-4}(N_j - q) + \sigma^{-4}\mathrm{tr}G_j^{-1} - \sigma^{-4}\mathrm{tr}\left\{\left[(Z_j'Z_j\Theta/\sigma^2)G_j^{-1}\right]G_j^{-1}\right\} \\
&= \sigma^{-4}(N_j - q) + \sigma^{-4}\mathrm{tr}G_j^{-1} - \sigma^{-4}\mathrm{tr}\left[(I_q - G_j^{-1})G_j^{-1}\right] \\
&= \sigma^{-4}(N_j - q) + \sigma^{-4}\mathrm{tr}G_j^{-1} - \sigma^{-4}\mathrm{tr}G_j^{-1} + \sigma^{-4}\mathrm{tr}G_j^{-2} \\
&= \sigma^{-4}(N_j - q) + \sigma^{-4}\mathrm{tr}G_j^{-2}. \qquad \text{(A.16)}
\end{aligned}$$

Differential formulas. As was stated in the previous section, the maximum likelihood estimates are obtained by minimizing the minus-log-likelihood function by the BFGS method, which uses the gradient of the function. To find the gradient, the differential notation of Magnus and Neudecker (1985, 1988) will be used. The key property of differentials is their relation with derivatives through the following equivalence: Let f be a vector or scalar function of a vector or scalar variable x, then

$$\frac{\partial f}{\partial x'} = A(x) \Leftrightarrow \mathrm{d}f = A(x)\,\mathrm{d}x.$$

The differential of a matrix is defined through the vector that stacks its columns: $\mathrm{vec}\,\mathrm{d}F = \mathrm{d}\,\mathrm{vec}F$. Note that the differential of a scalar, vector, or matrix is a scalar, vector, or matrix of the same size.

Some useful formulas are (Magnus & Neudecker, 1985, 1988):

$$\begin{aligned}
\mathrm{d}(c) &= 0 \\
\mathrm{d}(cg) &= c\,\mathrm{d}g \\
\mathrm{d}(g+h) &= \mathrm{d}g + \mathrm{d}h \\
\mathrm{d}(gh) &= (\mathrm{d}g)h + g\,\mathrm{d}h \\
\mathrm{d}(\log f) &= \frac{1}{f}\mathrm{d}f \\
\mathrm{d}(\det F) &= \det F\,\mathrm{tr}(F^{-1}\,\mathrm{d}F) \\
\mathrm{d}(\mathrm{tr}F) &= \mathrm{tr}\,\mathrm{d}F \\
\mathrm{d}(F^{-1}) &= -F^{-1}(\mathrm{d}F)F^{-1},
\end{aligned}$$

where c is a scalar, vector, or matrix constant, g and h may be scalars, vectors, or matrices (provided the expression is a valid expression), f is a scalar, and F is a matrix.

There is also a *chain rule*: If f is a function of x and g is a function of f, then (cf. Magnus & Neudecker, 1988, p. 91)

$$\frac{\partial g}{\partial x'} = \frac{\partial g}{\partial f'}\frac{\partial f}{\partial x'}.$$

This means the following for the differentials: If $\mathrm{d}g = A\mathrm{d}f$ and $\mathrm{d}f = B\mathrm{d}x$, then $\mathrm{d}g = AB\mathrm{d}x$, which illustrates that, informally speaking, the formulas for the differentials can be filled in sequentially. A similar formula holds for differentials of matrices. In the following, it will be clear how the chain rule can be applied.

The formulas above can be used to derive some important differentials:

$$\begin{aligned}
\mathrm{d}V_j &= \mathrm{d}(\sigma^2 I_{N_j} + Z_j\Theta Z_j') \\
&= (\mathrm{d}\sigma^2)I_{N_j} + Z_j(\mathrm{d}\Theta)Z_j' && \text{(A.17)} \\
\mathrm{d}\,\det V_j &= \det V_j\,\mathrm{tr}(V_j^{-1}\,\mathrm{d}V_j) && \text{(A.18)} \\
\mathrm{d}\,V_j^{-1} &= -V_j^{-1}(\mathrm{d}V_j)V_j^{-1} && \text{(A.19)} \\
\mathrm{d}\,\log\det V_j &= \frac{1}{\det V_j}\,\mathrm{d}\det V_j \\
&= \frac{1}{\det V_j}\,\det V_j\,\mathrm{tr}(V_j^{-1}\,\mathrm{d}V_j) \\
&= \mathrm{tr}(V_j^{-1}\,\mathrm{d}V_j). && \text{(A.20)}
\end{aligned}$$

Combining equation (A.19) with (A.17), we find that

$$\begin{aligned}
\mathrm{d}\,V_j^{-1} &= -V_j^{-1}[(\mathrm{d}\sigma^2)I_{N_j} + Z_j(\mathrm{d}\Theta)Z_j']V_j^{-1} \\
&= -V_j^{-2}\mathrm{d}\sigma^2 - V_j^{-1}Z_j(\mathrm{d}\Theta)Z_j'V_j^{-1}. && \text{(A.21)}
\end{aligned}$$

Consider a term of the form

$$\mathrm{d}T = \mathrm{tr}A\mathrm{d}\Theta,$$

where Θ is a symmetric $q \times q$ matrix. This term can be written as

$$\begin{aligned}
\mathrm{d}T &= \sum_{k=1}^{q}\sum_{l=1}^{q} A_{kl}\mathrm{d}\Theta_{lk} \\
&= \sum_{k=1}^{q}\sum_{l=1}^{k-1}(A_{kl} + A_{lk})\mathrm{d}\Theta_{kl} + \sum_{k=1}^{q} A_{kk}\mathrm{d}\Theta_{kk},
\end{aligned}$$

so

$$\begin{aligned}
\frac{\partial T}{\partial \Theta_{kl}} &= A_{kl} + A_{lk} && \text{(A.22)} \\
&= 2A_{kl}, \qquad \text{if } A \text{ is symmetric,} && \text{(A.23)}
\end{aligned}$$

and

$$\frac{\partial T}{\partial \Theta_{kk}} = A_{kk}, \qquad \text{(A.24)}$$

where $k \neq l$.

Similarly, consider a term of the form

$$\mathrm{d}S = [A(\mathrm{d}\Theta)B]_{kl},$$

where Θ is a symmetric $q \times q$ matrix, and A and B are matrices. This term can be written as

$$\begin{aligned}\mathrm{d}S &= \sum_{u=1}^{q}\sum_{v=1}^{q} A_{ku}(\mathrm{d}\Theta_{uv})B_{vl} \\ &= \sum_{u=1}^{q}\sum_{v=1}^{u-1}(A_{ku}B_{vl} + A_{kv}B_{ul})\mathrm{d}\Theta_{uv} + \sum_{u=1}^{q} A_{ku}B_{ul}\mathrm{d}\Theta_{uu},\end{aligned}$$

so

$$\frac{\partial S}{\partial \Theta_{uv}} = A_{ku}B_{vl} + A_{kv}B_{ul} \tag{A.25}$$

and

$$\frac{\partial S}{\partial \Theta_{uu}} = A_{ku}B_{ul}, \tag{A.26}$$

where $u \neq v$.

A.3 Computational formulas for the function and gradient

The formula (A.6) of the minus-log-likelihood function is computationally inefficient, because a matrix of size N_j has to be inverted, and its determinant calculated. Therefore, in this section a computationally efficient formula will be derived, based on Longford (1987), and using formulas from the previous section. Along the same lines, computationally efficient formulas for the derivatives of this function with respect to the parameters will also be derived.

Combining (A.6), (A.9), and (A.11), we find the following formula for L:

$$\begin{aligned}
L &= \frac{N}{2}\log 2\pi + \frac{1}{2}\sum_{j=1}^{J}\log[(\sigma^2)^{N_j}\det G_j] \\
&\quad + \frac{1}{2}\sum_{j=1}^{J}(y_j - X_j\gamma)'[\sigma^{-2}I_{N_j} - \sigma^{-4}Z_j\Theta G_j^{-1}Z_j'](y_j - X_j\gamma) \\
&= \frac{N}{2}\log 2\pi + \frac{N}{2}\log(\sigma^2) + \frac{1}{2}\sum_{j=1}^{J}\log\det G_j \\
&\quad + \frac{1}{2}\sigma^{-2}\sum_{j=1}^{J}(y_j - X_j\gamma)'(y_j - X_j\gamma) \\
&\quad - \frac{1}{2}\sigma^{-4}\sum_{j=1}^{J}(y_j - X_j\gamma)'Z_j\Theta G_j^{-1}Z_j'(y_j - X_j\gamma) \\
&= \frac{N}{2}\log 2\pi + \frac{N}{2}\log(\sigma^2) + \frac{1}{2}\sum_{j=1}^{J}\log\det G_j \\
&\quad + \frac{1}{2}\sigma^{-2}\left[\left(\sum_{j=1}^{J}y_j'y_j\right) - 2\gamma'\left(\sum_{j=1}^{J}X_j'y_j\right) + \gamma'\left(\sum_{j=1}^{J}X_j'X_j\right)\gamma\right] \\
&\quad - \frac{1}{2}\sigma^{-4}\sum_{j=1}^{J}(Z_j'y_j - Z_j'X_j\gamma)'\Theta G_j^{-1}(Z_j'y_j - Z_j'X_j\gamma). \qquad \text{(A.27)}
\end{aligned}$$

Formula (A.27) is a computationally efficient formula, and this is the formula that is implemented in the program.

To find the gradient of L, we start with the differential of L:

$$\begin{aligned}
\mathrm{d}L &= \frac{1}{2}\sum_{j=1}^{J}\mathrm{d}\,\log\,\det V_j \\
&\quad + \frac{1}{2}\sum_{j=1}^{J}2(y_j - X_j\gamma)'V_j^{-1}(-X_j\mathrm{d}\gamma) \\
&\quad + \frac{1}{2}\sum_{j=1}^{J}(y_j - X_j\gamma)'\mathrm{d}(V_j^{-1})(y_j - X_j\gamma). \qquad \text{(A.28)}
\end{aligned}$$

Combining (A.28) with (A.20), (A.21) and (A.17), we find

$$\begin{aligned}
\mathrm{d}L &= \frac{1}{2}\sum_{j=1}^{J}\mathrm{tr}(V_j^{-1}\mathrm{d}V_j) - \sum_{j=1}^{J}(y_j - X_j\gamma)'V_j^{-1}X_j\mathrm{d}\gamma \\
&\quad - \frac{1}{2}\sum_{j=1}^{J}(y_j - X_j\gamma)'[V_j^{-2}\mathrm{d}\sigma^2 + V_j^{-1}Z_j(\mathrm{d}\Theta)Z_j'V_j^{-1}](y_j - X_j\gamma)
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{2}\sum_{j=1}^{J} \mathrm{tr}\{V_j^{-1}[(\mathrm{d}\sigma^2) I_{N_j} + Z_j(\mathrm{d}\Theta)Z_j']\} \\
&\quad - \sum_{j=1}^{J}(y_j - X_j\gamma)'V_j^{-1}X_j\mathrm{d}\gamma \\
&\quad - \frac{1}{2}\sum_{j=1}^{J}(y_j - X_j\gamma)'V_j^{-2}(\mathrm{d}\sigma^2)(y_j - X_j\gamma) \\
&\quad - \frac{1}{2}\sum_{j=1}^{J}(y_j - X_j\gamma)'V_j^{-1}Z_j(\mathrm{d}\Theta)Z_j'V_j^{-1}(y_j - X_j\gamma) \\
&= \left(\frac{1}{2}\sum_{j=1}^{J}\mathrm{tr}V_j^{-1}\right)\mathrm{d}\sigma^2 + \frac{1}{2}\mathrm{tr}[V_j^{-1}Z_j(\mathrm{d}\Theta)Z_j'] \\
&\quad - \sum_{j=1}^{J}(y_j - X_j\gamma)'V_j^{-1}X_j\mathrm{d}\gamma \\
&\quad - \frac{1}{2}\sum_{j=1}^{J}\left[(y_j - X_j\gamma)'V_j^{-2}(y_j - X_j\gamma)\right]\mathrm{d}\sigma^2 \\
&\quad - \frac{1}{2}\sum_{j=1}^{J}\mathrm{tr}\left\{[(y_j - X_j\gamma)'V_j^{-1}Z_j](\mathrm{d}\Theta)[Z_j'V_j^{-1}(y_j - X_j\gamma)]\right\} \\
&= \left(\frac{1}{2}\sum_{j=1}^{J}\mathrm{tr}V_j^{-1}\right)\mathrm{d}\sigma^2 + \frac{1}{2}\mathrm{tr}(Z_j'V_j^{-1}Z_j\mathrm{d}\Theta) \\
&\quad - \left[\sum_{j=1}^{J}(y_j - X_j\gamma)'V_j^{-1}X_j\right]\mathrm{d}\gamma \\
&\quad - \left[\frac{1}{2}\sum_{j=1}^{J}(y_j - X_j\gamma)'V_j^{-2}(y_j - X_j\gamma)\right]\mathrm{d}\sigma^2 \\
&\quad - \frac{1}{2}\sum_{j=1}^{J}\mathrm{tr}\left(\left\{\left[Z_j'V_j^{-1}(y_j - X_j\gamma)\right]\left[Z_j'V_j^{-1}(y_j - X_j\gamma)\right]'\right\}\mathrm{d}\Theta\right).
\end{aligned}
$$

So

$$
\frac{\partial L}{\partial \gamma'} = -\sum_{j=1}^{J}(y_j - X_j\gamma)'V_j^{-1}X_j \tag{A.29}
$$

and

$$
\frac{\partial L}{\partial \sigma^2} = \frac{1}{2}\sum_{j=1}^{J}\mathrm{tr}V_j^{-1} - \frac{1}{2}\sum_{j=1}^{J}(y_j - X_j\gamma)'V_j^{-2}(y_j - X_j\gamma), \tag{A.30}
$$

and, using (A.23) and (A.24),

$$
\begin{aligned}
\frac{\partial L}{\partial \Theta_{kl}} &= \sum_{j=1}^{J}(Z_j'V_j^{-1}Z_j)_{kl} \\
&\quad - \sum_{j=1}^{J}\left\{\left[Z_j'V_j^{-1}(y_j - X_j\gamma)\right]\left[Z_j'V_j^{-1}(y_j - X_j\gamma)\right]\right\}_{kl}
\end{aligned}
$$

$$= \sum_{j=1}^{J} (Z_j' V_j^{-1} Z_j)_{kl} - \sum_{j=1}^{J} \left[Z_j' V_j^{-1} (y_j - X_j \gamma) \right]_k \left[Z_j' V_j^{-1} (y_j - X_j \gamma) \right]_l \tag{A.31}$$

and

$$\frac{\partial L}{\partial \Theta_{kk}} = \frac{1}{2} \sum_{j=1}^{J} (Z_j' V_j^{-1} Z_j)_{kk} - \frac{1}{2} \sum_{j=1}^{J} \left[Z_j' V_j^{-1} (y_j - X_j \gamma) \right]_k^2 . \tag{A.32}$$

Now, using (A.9), (A.12), (A.13), and (A.15), we find computationally more efficient formulas for the derivatives:

$$\begin{aligned}
\frac{\partial L}{\partial \gamma} &= - \sum_{j=1}^{J} X_j' V_j^{-1} (y_j - X_j \gamma) \\
&= - \sum_{j=1}^{J} X_j' (\sigma^{-2} I_{N_j} - \sigma^{-4} Z_j \Theta G_j^{-1} Z_j') (y_j - X_j \gamma) \\
&= -\sigma^{-2} \left[\left(\sum_{j=1}^{J} X_j' y_j \right) - \left(\sum_{j=1}^{J} X_j' X_j \right) \gamma \right] \\
&\quad + \sigma^{-4} \sum_{j=1}^{J} X_j' Z_j \Theta G_j^{-1} (Z_j' y_j - Z_j' X_j \gamma), \\
\frac{\partial L}{\partial \sigma^2} &= \frac{1}{2} \sum_{j=1}^{J} \mathrm{tr} V_j^{-1} - \frac{1}{2} \sum_{j=1}^{J} (y_j - X_j \gamma)' V_j^{-2} (y_j - X_j \gamma) \\
&= \frac{1}{2} \sigma^{-2} (N - Jq) + \frac{1}{2} \sigma^{-2} \sum_{j=1}^{J} \mathrm{tr} G_j^{-1} \\
&\quad - \frac{1}{2} \sum_{j=1}^{J} (y_j - X_j \gamma)' (\sigma^{-2} I_{N_j} - \sigma^{-4} Z_j \Theta G_j^{-1} Z_j') V_j^{-1} (y_j - X_j \gamma) \\
&= \frac{1}{2} \sigma^{-2} (N - Jq) + \frac{1}{2} \sigma^{-2} \sum_{j=1}^{J} \mathrm{tr} G_j^{-1} \\
&\quad - \frac{1}{2} \sigma^{-2} \sum_{j=1}^{J} (y_j - X_j \gamma)' V_j^{-1} (y_j - X_j \gamma) \\
&\quad + \frac{1}{2} \sigma^{-4} \sum_{j=1}^{J} (y_j - X_j \gamma)' Z_j \Theta G_j^{-1} Z_j' V_j^{-1} (y_j - X_j \gamma) \\
&= \frac{1}{2} \sigma^{-2} (N - Jq) + \frac{1}{2} \sigma^{-2} \sum_{j=1}^{J} \mathrm{tr} G_j^{-1} \\
&\quad - \frac{1}{2} \sigma^{-2} \sum_{j=1}^{J} (y_j - X_j \gamma)' (\sigma^{-2} I_{N_j} - \sigma^{-4} Z_j \Theta G_j^{-1} Z_j') (y_j - X_j \gamma) \\
&\quad + \frac{1}{2} \sigma^{-4} \sum_{j=1}^{J} (y_j - X_j \gamma)' Z_j \Theta G_j^{-1} (\sigma^{-2} G_j^{-1} Z_j') (y_j - X_j \gamma)
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{2}\sigma^{-2}(N-nq) + \frac{1}{2}\sigma^{-2}\sum_{j=1}^{J}\mathrm{tr}G_j^{-1} \\
&\quad - \frac{1}{2}\sigma^{-4}\left[\left(\sum_{j=1}^{J} y_j'y_j\right) - 2\gamma'\left(\sum_{j=1}^{J} X_j'y_j\right) + \gamma'\left(\sum_{j=1}^{J} X_j'X_j\right)\gamma\right] \\
&\quad + \frac{1}{2}\sigma^{-6}\sum_{j=1}^{J}(Z_j'y_j - Z_j'X_j\gamma)'\Theta G_j^{-1}(Z_j'y_j - Z_j'X_j\gamma) \\
&\quad + \frac{1}{2}\sigma^{-6}\sum_{j=1}^{J}(Z_j'y_j - Z_j'X_j\gamma)'\Theta G_j^{-2}(Z_j'y_j - Z_j'X_j\gamma) \\
&= \frac{1}{2}\sigma^{-2}(N-nq) + \frac{1}{2}\sigma^{-2}\sum_{j=1}^{J}\mathrm{tr}G_j^{-1} \\
&\quad - \frac{1}{2}\sigma^{-4}\left[\left(\sum_{j=1}^{J} y_j'y_j\right) - 2\gamma'\left(\sum_{j=1}^{J} X_j'y_j\right) + \gamma'\left(\sum_{j=1}^{J} X_j'X_j\right)\gamma\right] \\
&\quad + \frac{1}{2}\sigma^{-6}\sum_{j=1}^{J}(Z_j'y_j - Z_j'X_j\gamma)'\Theta(I_q + G_j^{-1})G_j^{-1}(Z_j'y_j - Z_j'X_j\gamma), \\
\frac{\partial L}{\partial \Theta_{kl}} &= \sum_{j=1}^{J}(Z_j'V_j^{-1}Z_j)_{kl} \\
&\quad - \sum_{j=1}^{J}[Z_j'V_j^{-1}(y_j - X_j\gamma)]_k[Z_j'V_j^{-1}(y_j - X_j\gamma)]_l \\
&= \sum_{j=1}^{J}(\sigma^{-2}G_j^{-1}Z_j'Z_j)_{kl} \\
&\quad - \sum_{j=1}^{J}[\sigma^{-2}G_j^{-1}(Z_j'y_j - Z_j'X_j\gamma)]_k[\sigma^{-2}G_j^{-1}(Z_j'y_j - Z_j'X_j\gamma)]_l,
\end{aligned}
$$

and

$$
\frac{\partial L}{\partial \Theta_{kk}} = \frac{1}{2}\sum_{j=1}^{J}(\sigma^{-2}G_j^{-1}Z_j'Z_j)_{kk} - \frac{1}{2}\sum_{j=1}^{J}[\sigma^{-2}G_j^{-1}(Z_j'y_j - Z_j'X_j\gamma)]_k^2.
$$

These formulas are implemented in the program. Note (cf. Bryk & Raudenbush, 1992, p. 239) that the function and the derivatives depend on the data only through the terms $\sum_{j=1}^{J} y_j'y_j$, $\sum_{j=1}^{J} X_j'y_j$, $\sum_{j=1}^{J} X_j'X_j$, $Z_j'y_j$, $Z_j'X_j$, and $Z_j'Z_j$ (through G_j), the first of which is a scalar, the second a p-vector, and the third a symmetric $p \times p$ matrix. The last three are a q-vector, a $q \times p$ matrix, and a symmetric $q \times q$ matrix, for each Level-2 unit. The symmetric matrices may be stored linearly, thereby saving additional memory.

A.4 The asymptotic covariance matrix of the estimators

The asymptotic distribution of the maximum likelihood estimators, under appropriate general conditions is given by (see Magnus, 1978)

$$
\sqrt{N}\left(\hat{\theta}_{\mathrm{ML}} - \theta\right) \xrightarrow{\mathcal{L}} N\left[0, \lim_{N\to\infty}\left(\frac{I(\theta)}{N}\right)^{-1}\right], \tag{A.33}
$$

where N is the sample size,

$$I(\theta) = E\left(\frac{\partial^2 L}{\partial\theta\,\partial\theta'}\right), \tag{A.34}$$

and L is the minus-log-likelihood function. Therefore, the asymptotic covariance matrix of the estimators is derived from the matrix of second derivatives of L (the Hessian matrix).

From (A.29) we have

$$\frac{\partial L}{\partial\gamma} = -\sum_{j=1}^{J} X_j' V_j^{-1}(y_j - X_j\gamma).$$

Thus,

$$\begin{aligned}
\mathrm{d}\left(\frac{\partial L}{\partial\gamma}\right) &= -\sum_{j=1}^{J} X_j'(\mathrm{d}V_j^{-1})(y_j - X_j\gamma) + \sum_{j=1}^{J} X_j' V_j^{-1} X_j\,\mathrm{d}\gamma \\
&= \sum_{j=1}^{J} X_j'[V_j^{-2}\mathrm{d}\sigma^2 + V_j^{-1} Z_j(\mathrm{d}\Theta) Z_j' V_j^{-1}](y_j - X_j\gamma) \\
&\quad + \sum_{j=1}^{J} X_j' V_j^{-1} X_j\,\mathrm{d}\gamma \qquad \text{(using (A.21))} \\
&= \left[\sum_{j=1}^{J} X_j' V_j^{-2}(y_j - X_j\gamma)\right]\mathrm{d}\sigma^2 \\
&\quad + \sum_{j=1}^{J}(X_j' V_j^{-1} Z_j)(\mathrm{d}\Theta)[Z_j' V_j^{-1}(y_j - X_j\gamma)] \\
&\quad + \left(\sum_{j=1}^{J} X_j' V_j^{-1} X_j\right)\mathrm{d}\gamma.
\end{aligned} \tag{A.35}$$

Therefore,

$$\frac{\partial^2 L}{\partial\gamma\,\partial\gamma'} = \sum_{j=1}^{J} X_j' V_j^{-1} X_j, \tag{A.36}$$

$$\frac{\partial^2 L}{\partial\gamma\,\partial\sigma^2} = \sum_{j=1}^{J} X_j' V_j^{-2}(y_j - X_j\gamma), \tag{A.37}$$

and, using (A.25) and (A.26),

$$\begin{aligned}
\frac{\partial^2 L}{\partial\gamma_k\,\partial\Theta_{uv}} = \sum_{j=1}^{J} \Big\{ &(X_j' V_j^{-1} Z_j)_{ku}[Z_j' V_j^{-1}(y_j - X_j\gamma)]_v \\
&+ (X_j' V_j^{-1} Z_j)_{kv}[Z_j' V_j^{-1}(y_j - X_j\gamma)]_u \Big\},
\end{aligned} \tag{A.38}$$

$$\frac{\partial^2 L}{\partial\gamma_k\,\partial\Theta_{uu}} = \sum_{j=1}^{J}(X_j' V_j^{-1} Z_j)_{ku}[Z_j' V_j^{-1}(y_j - X_j\gamma)]_u. \tag{A.39}$$

From (A.30), we have

$$\frac{\partial L}{\partial\sigma^2} = \frac{1}{2}\sum_{j=1}^{J}\mathrm{tr}(V_j^{-1}) - \frac{1}{2}\sum_{j=1}^{J}(y_j - X_j\gamma)' V_j^{-2}(y_j - X_j\gamma).$$

Thus,

$$
\begin{aligned}
\mathrm{d}\left(\frac{\partial L}{\partial \sigma^2}\right) &= \frac{1}{2}\sum_{j=1}^{J}\mathrm{tr}(\mathrm{d}V_j^{-1}) - \frac{1}{2}\sum_{j=1}^{J}(y_j - X_j\gamma)'(\mathrm{d}V_j^{-2})(y_j - X_j\gamma) \\
&\quad + \sum_{j=1}^{J}(y_j - X_j\gamma)'V_j^{-2}X_j\,\mathrm{d}\gamma \\
&= -\frac{1}{2}\sum_{j=1}^{J}\mathrm{tr}[V_j^{-2}\mathrm{d}\sigma^2 + V_j^{-1}Z_j(\mathrm{d}\Theta)Z_j'V_j^{-1}] \\
&\quad - \frac{1}{2}\sum_{j=1}^{J}(y_j - X_j\gamma)'(\mathrm{d}V_j^{-2})(y_j - X_j\gamma) \\
&\quad + \sum_{j=1}^{J}(y_j - X_j\gamma)'V_j^{-2}X_j\,\mathrm{d}\gamma \quad \text{(using (A.21))} \\
&= \left(-\frac{1}{2}\sum_{j=1}^{J}\mathrm{tr}V_j^{-2}\right)\mathrm{d}\sigma^2 - \frac{1}{2}\sum_{j=1}^{J}\mathrm{tr}[(V_j^{-1}Z_j)(\mathrm{d}\Theta)(Z_j'V_j^{-1})] \\
&\quad - \frac{1}{2}\sum_{j=1}^{J}(y_j - X_j\gamma)'[(\mathrm{d}V_j^{-1})V_j^{-1} + V_j^{-1}\,\mathrm{d}V_j^{-1}](y_j - X_j\gamma) \\
&\quad + \sum_{j=1}^{J}(y_j - X_j\gamma)'V_j^{-2}X_j\,\mathrm{d}\gamma \\
&= \left(-\frac{1}{2}\sum_{j=1}^{J}\mathrm{tr}V_j^{-2}\right)\mathrm{d}\sigma^2 - \frac{1}{2}\sum_{j=1}^{J}\mathrm{tr}[(Z_j'V_j^{-2}Z_j)\mathrm{d}\Theta] \\
&\quad - \sum_{j=1}^{J}(y_j - X_j\gamma)'V_j^{-1}(\mathrm{d}V_j^{-1})(y_j - X_j\gamma) \\
&\quad + \sum_{j=1}^{J}(y_j - X_j\gamma)'V_j^{-2}X_j\,\mathrm{d}\gamma \\
&= \left(-\frac{1}{2}\sum_{j=1}^{J}\mathrm{tr}V_j^{-2}\right)\mathrm{d}\sigma^2 - \frac{1}{2}\sum_{j=1}^{J}\mathrm{tr}[(Z_j'V_j^{-2}Z_j)\mathrm{d}\Theta] \\
&\quad + \sum_{j=1}^{J}(y_j - X_j\gamma)'V_j^{-1}[V_j^{-2}\mathrm{d}\sigma^2 + V_j^{-1}Z_j(\mathrm{d}\Theta)Z_j'V_j^{-1}](y_j - X_j\gamma) \\
&\quad + \sum_{j=1}^{J}(y_j - X_j\gamma)'V_j^{-2}X_j\,\mathrm{d}\gamma
\end{aligned}
$$

$$
\begin{aligned}
&= \left(-\frac{1}{2}\sum_{j=1}^{J}\mathrm{tr}V_j^{-2}\right)\mathrm{d}\sigma^2 - \frac{1}{2}\sum_{j=1}^{J}\mathrm{tr}[(Z_j'V_j^{-2}Z_j)\mathrm{d}\Theta] \\
&\quad + \left[\sum_{j=1}^{J}(y_j - X_j\gamma)'V_j^{-3}(y_j - X_j\gamma)\right]\mathrm{d}\sigma^2 \\
&\quad + \sum_{j=1}^{J}[(y_j - X_j\gamma)'V_j^{-2}Z_j](\mathrm{d}\Theta)[Z_j'V_j^{-1}(y_j - X_j\gamma)] \\
&\quad + \sum_{j=1}^{J}(y_j - X_j\gamma)'V_j^{-2}X_j\,\mathrm{d}\gamma. \\
&= \left(-\frac{1}{2}\sum_{j=1}^{J}\mathrm{tr}V_j^{-2}\right)\mathrm{d}\sigma^2 - \frac{1}{2}\sum_{j=1}^{J}\mathrm{tr}[(Z_j'V_j^{-2}Z_j)\mathrm{d}\Theta] \\
&\quad + \left[\sum_{j=1}^{J}(y_j - X_j\gamma)'V_j^{-3}(y_j - X_j\gamma)\right]\mathrm{d}\sigma^2 \\
&\quad + \sum_{j=1}^{J}\mathrm{tr}\left\{\left[Z_j'V_j^{-1}(y_j - X_j\gamma)(y_j - X_j\gamma)'V_j^{-2}Z_j\right]\mathrm{d}\Theta\right\} \\
&\quad + \sum_{j=1}^{J}(y_j - X_j\gamma)'V_j^{-2}X_j\,\mathrm{d}\gamma. \qquad \text{(A.40)}
\end{aligned}
$$

Therefore,

$$
\frac{\partial^2 L}{\partial\sigma^2\,\partial\sigma^2} = -\frac{1}{2}\sum_{j=1}^{J}\mathrm{tr}V_j^{-2} + \sum_{j=1}^{J}(y_j - X_j\gamma)'V_j^{-3}(y_j - X_j\gamma), \qquad \text{(A.41)}
$$

and, using (A.22), (A.23) and (A.24),

$$
\begin{aligned}
\frac{\partial^2 L}{\partial\sigma^2\,\partial\Theta_{kl}} &= -\sum_{j=1}^{J}(Z_j'V_j^{-2}Z_j)_{kl} \\
&\quad + \sum_{j=1}^{J}\left\{\left[Z_j'V_j^{-1}(y_j - X_j\gamma)(y_j - X_j\gamma)'V_j^{-2}Z_j\right]_{kl}\right. \\
&\qquad \left. + \left[Z_j'V_j^{-1}(y_j - X_j\gamma)(y_j - X_j\gamma)'V_j^{-2}Z_j\right]_{lk}\right\}, \qquad \text{(A.42)} \\
\frac{\partial^2 L}{\partial\sigma^2\,\partial\Theta_{kk}} &= -\frac{1}{2}\sum_{j=1}^{J}(Z_j'V_j^{-2}Z_j)_{kk} \\
&\quad + \sum_{j=1}^{J}\left[Z_j'V_j^{-1}(y_j - X_j\gamma)(y_j - X_j\gamma)'V_j^{-2}Z_j\right]_{kk} \qquad \text{(A.43)}
\end{aligned}
$$

From (A.31) we have

$$
\frac{\partial L}{\partial\Theta_{kl}} = \sum_{j=1}^{J}(Z_j'V_j^{-1}Z_j)_{kl} - \sum_{j=1}^{J}\left[Z_j'V_j^{-1}(y_j - X_j\gamma)\right]_k\left[Z_j'V_j^{-1}(y_j - X_j\gamma)\right]_l.
$$

Thus,

$$
\begin{aligned}
\mathrm{d}\left(\frac{\partial L}{\partial \Theta_{kl}}\right) &= \sum_{j=1}^{J} [Z_j'(\mathrm{d}V_j^{-1})Z_j]_{kl} \\
&\quad - \sum_{j=1}^{J} \left[Z_j'(\mathrm{d}V_j^{-1})(y_j - X_j\gamma)\right]_k \left[Z_j'V_j^{-1}(y_j - X_j\gamma)\right]_l \\
&\quad - \sum_{j=1}^{J} \left[Z_j'V_j^{-1}(y_j - X_j\gamma)\right]_k \left[Z_j'(\mathrm{d}V_j^{-1})(y_j - X_j\gamma)\right]_l \\
&\quad + \sum_{j=1}^{J} \left(Z_j'V_j^{-1}X_j\,\mathrm{d}\gamma\right)_k \left[Z_j'V_j^{-1}(y_j - X_j\gamma)\right]_l \\
&\quad + \sum_{j=1}^{J} \left[Z_j'V_j^{-1}(y_j - X_j\gamma)\right]_k \left(Z_j'V_j^{-1}X_j\,\mathrm{d}\gamma\right)_l \\
&= -\sum_{j=1}^{J} \left\{Z_j'[V_j^{-2}\mathrm{d}\sigma^2 + V_j^{-1}Z_j(\mathrm{d}\Theta)Z_j'V_j^{-1}]Z_j\right\}_{kl} \\
&\quad + \sum_{j=1}^{J} \left\{Z_j'[V_j^{-2}\mathrm{d}\sigma^2 + V_j^{-1}Z_j(\mathrm{d}\Theta)Z_j'V_j^{-1}](y_j - X_j\gamma)\right\}_k \\
&\qquad\quad \times \left[Z_j'V_j^{-1}(y_j - X_j\gamma)\right]_l \\
&\quad + \sum_{j=1}^{J} \left[Z_j'V_j^{-1}(y_j - X_j\gamma)\right]_k \\
&\qquad\quad \times \left\{Z_j'[V_j^{-2}\mathrm{d}\sigma^2 + V_j^{-1}Z_j(\mathrm{d}\Theta)Z_j'V_j^{-1}](y_j - X_j\gamma)\right\}_l \\
&\quad + \sum_{j=1}^{J} \left(Z_j'V_j^{-1}X_j\,\mathrm{d}\gamma\right)_k \left[Z_j'V_j^{-1}(y_j - X_j\gamma)\right]_l \\
&\quad + \sum_{j=1}^{J} \left[Z_j'V_j^{-1}(y_j - X_j\gamma)\right]_k \left(Z_j'V_j^{-1}X_j\,\mathrm{d}\gamma\right)_l \quad \text{(using (A.21))}
\end{aligned}
$$

$$
\begin{aligned}
&= \left[-\sum_{j=1}^{J} (Z_j' V_j^{-2} Z_j)_{kl} \right] \mathrm{d}\sigma^2 - \sum_{j=1}^{J} \left[(Z_j' V_j^{-1} Z_j)(\mathrm{d}\Theta)(Z_j' V_j^{-1} Z_j) \right]_{kl} \\
&\quad + \left\{ \sum_{j=1}^{J} \left[Z_j' V_j^{-2} (y_j - X_j\gamma) \right]_k \left[Z_j' V_j^{-1} (y_j - X_j\gamma) \right]_l \right\} \mathrm{d}\sigma^2 \\
&\quad + \sum_{j=1}^{J} \Big\{ (Z_j' V_j^{-1} Z_j)(\mathrm{d}\Theta) \\
&\qquad\quad \times \left[Z_j' V_j^{-1} (y_j - X_j\gamma)(y_j - X_j\gamma)' V_j^{-1} Z_j \right] \Big\}_{kl} \\
&\quad + \left\{ \sum_{j=1}^{J} \left[Z_j' V_j^{-1} (y_j - X_j\gamma) \right]_k \left[Z_j' V_j^{-2} (y_j - X_j\gamma) \right]_l \right\} \mathrm{d}\sigma^2 \\
&\quad + \sum_{j=1}^{J} \Big\{ \left[Z_j' V_j^{-1} (y_j - X_j\gamma)(y_j - X_j\gamma)' V_j^{-1} Z_j \right] \\
&\qquad\quad \times (\mathrm{d}\Theta)(Z_j' V_j^{-1} Z_j) \Big\}_{kl} \\
&\quad + \sum_{j=1}^{J} \left(Z_j' V_j^{-1} X_j \, \mathrm{d}\gamma \right)_k \left[Z_j' V_j^{-1} (y_j - X_j\gamma) \right]_l \\
&\quad + \sum_{j=1}^{J} \left[Z_j' V_j^{-1} (y_j - X_j\gamma) \right]_k \left(Z_j' V_j^{-1} X_j \, \mathrm{d}\gamma \right)_l
\end{aligned} \tag{A.44}
$$

Combining (A.44) with (A.25) and (A.26), the partial derivatives are found:

$$
\begin{aligned}
\frac{\partial^2 L}{\partial\Theta_{kl}\,\partial\Theta_{uv}} &= -\sum_{j=1}^{J} \left[(Z_j' V_j^{-1} Z_j)_{ku} (Z_j' V_j^{-1} Z_j)_{vl} + (Z_j' V_j^{-1} Z_j)_{kv} (Z_j' V_j^{-1} Z_j)_{ul} \right] \\
&\quad + \sum_{j=1}^{J} \Big\{ (Z_j' V_j^{-1} Z_j)_{ku} \left[Z_j' V_j^{-1} (y_j - X_j\gamma)(y_j - X_j\gamma)' V_j^{-1} Z_j \right]_{vl} \\
&\qquad + (Z_j' V_j^{-1} Z_j)_{kv} \left[Z_j' V_j^{-1} (y_j - X_j\gamma)(y_j - X_j\gamma)' V_j^{-1} Z_j \right]_{ul} \Big\} \\
&\quad + \sum_{j=1}^{J} \Big\{ \left[Z_j' V_j^{-1} (y_j - X_j\gamma)(y_j - X_j\gamma)' V_j^{-1} Z_j \right]_{ku} (Z_j' V_j^{-1} Z_j)_{vl} \\
&\qquad + \left[Z_j' V_j^{-1} (y_j - X_j\gamma)(y_j - X_j\gamma)' V_j^{-1} Z_j \right]_{kv} (Z_j' V_j^{-1} Z_j)_{ul} \Big\} \\
&= -\sum_{j=1}^{J} \left[(Z_j' V_j^{-1} Z_j)_{ku} (Z_j' V_j^{-1} Z_j)_{vl} + (Z_j' V_j^{-1} Z_j)_{kv} (Z_j' V_j^{-1} Z_j)_{ul} \right] \\
&\quad + \sum_{j=1}^{J} \Big\{ (Z_j' V_j^{-1} Z_j)_{ku} \left[Z_j' V_j^{-1} (y_j - X_j\gamma)(y_j - X_j\gamma)' V_j^{-1} Z_j \right]_{vl} \\
&\qquad + (Z_j' V_j^{-1} Z_j)_{kv} \left[Z_j' V_j^{-1} (y_j - X_j\gamma)(y_j - X_j\gamma)' V_j^{-1} Z_j \right]_{ul} \\
&\qquad + (Z_j' V_j^{-1} Z_j)_{ul} \left[Z_j' V_j^{-1} (y_j - X_j\gamma)(y_j - X_j\gamma)' V_j^{-1} Z_j \right]_{kv} \\
&\qquad + (Z_j' V_j^{-1} Z_j)_{vl} \left[Z_j' V_j^{-1} (y_j - X_j\gamma)(y_j - X_j\gamma)' V_j^{-1} Z_j \right]_{ku} \Big\}
\end{aligned} \tag{A.45}
$$

and

$$\begin{aligned}\frac{\partial^2 L}{\partial\Theta_{kl}\,\partial\Theta_{uu}} &= -\sum_{j=1}^{J}\left[(Z_j'V_j^{-1}Z_j)_{ku}(Z_j'V_j^{-1}Z_j)_{ul}\right]\\ &\quad+\sum_{j=1}^{J}\left\{(Z_j'V_j^{-1}Z_j)_{ku}\left[Z_j'V_j^{-1}(y_j-X_j\gamma)(y_j-X_j\gamma)'V_j^{-1}Z_j\right]_{ul}\right\}\\ &\quad+\sum_{j=1}^{J}\left\{\left[Z_j'V_j^{-1}(y_j-X_j\gamma)(y_j-X_j\gamma)'V_j^{-1}Z_j\right]_{ku}(Z_j'V_j^{-1}Z_j)_{ul}\right\}\\ &= -\sum_{j=1}^{J}\left[(Z_j'V_j^{-1}Z_j)_{ku}(Z_j'V_j^{-1}Z_j)_{ul}\right]\\ &\quad+\sum_{j=1}^{J}\left\{(Z_j'V_j^{-1}Z_j)_{ku}\left[Z_j'V_j^{-1}(y_j-X_j\gamma)(y_j-X_j\gamma)'V_j^{-1}Z_j\right]_{ul}\right.\\ &\qquad\left.+\left[Z_j'V_j^{-1}(y_j-X_j\gamma)(y_j-X_j\gamma)'V_j^{-1}Z_j\right]_{ku}(Z_j'V_j^{-1}Z_j)_{ul}\right\}.\end{aligned}\tag{A.46}$$

Analogously, from (A.32) we have

$$\frac{\partial L}{\partial\Theta_{kk}} = \frac{1}{2}\sum_{j=1}^{J}(Z_j'V_j^{-1}Z_j)_{kk} - \frac{1}{2}\sum_{j=1}^{J}\left[Z_j'V_j^{-1}(y_j-X_j\gamma)\right]_k^2,$$

and

$$\begin{aligned}\mathrm{d}\left(\frac{\partial L}{\partial\Theta_{kk}}\right) &= \left[-\frac{1}{2}\sum_{j=1}^{J}(Z_j'V_j^{-2}Z_j)_{kk}\right]\mathrm{d}\sigma^2\\ &\quad-\frac{1}{2}\sum_{j=1}^{J}\left[(Z_j'V_j^{-1}Z_j)(\mathrm{d}\Theta)(Z_j'V_j^{-1}Z_j)\right]_{kk}\\ &\quad+\left\{\frac{1}{2}\sum_{j=1}^{J}\left[Z_j'V_j^{-2}(y_j-X_j\gamma)\right]_k\left[Z_j'V_j^{-1}(y_j-X_j\gamma)\right]_k\right\}\mathrm{d}\sigma^2\\ &\quad+\frac{1}{2}\sum_{j=1}^{J}\left\{(Z_j'V_j^{-1}Z_j)(\mathrm{d}\Theta)\right.\\ &\qquad\left.\times\left[Z_j'V_j^{-1}(y_j-X_j\gamma)(y_j-X_j\gamma)'V_j^{-1}Z_j\right]\right\}_{kk}\\ &\quad+\left\{\frac{1}{2}\sum_{j=1}^{J}\left[Z_j'V_j^{-1}(y_j-X_j\gamma)\right]_k\left[Z_j'V_j^{-2}(y_j-X_j\gamma)\right]_k\right\}\mathrm{d}\sigma^2\\ &\quad+\frac{1}{2}\sum_{j=1}^{J}\left\{\left[Z_j'V_j^{-1}(y_j-X_j\gamma)(y_j-X_j\gamma)'V_j^{-1}Z_j\right]\right.\\ &\qquad\left.\times(\mathrm{d}\Theta)(Z_j'V_j^{-1}Z_j)\right\}_{kk}\\ &\quad+\frac{1}{2}\sum_{j=1}^{J}(Z_j'V_j^{-1}X_j\,\mathrm{d}\gamma)_k\left[Z_j'V_j^{-1}(y_j-X_j\gamma)\right]_k\\ &\quad+\frac{1}{2}\sum_{j=1}^{J}\left[Z_j'V_j^{-1}(y_j-X_j\gamma)\right]_k(Z_j'V_j^{-1}X_j\,\mathrm{d}\gamma)_k.\end{aligned}\tag{A.47}$$

Combining (A.47) with (A.26) we have

$$\begin{aligned}
\frac{\partial^2 L}{\partial\Theta_{kk}\,\partial\Theta_{uu}} &= -\frac{1}{2}\sum_{j=1}^{J}(Z_j'V_j^{-1}Z_j)_{ku}^2 \\
&\quad + \frac{1}{2}\sum_{j=1}^{J}(Z_j'V_j^{-1}Z_j)_{ku} \\
&\qquad \times \left[Z_j'V_j^{-1}(y_j - X_j\gamma)(y_j - X_j\gamma)'V_j^{-1}Z_j\right]_{uk} \\
&\quad + \frac{1}{2}\sum_{j=1}^{J}\left[Z_j'V_j^{-1}(y_j - X_j\gamma)(y_j - X_j\gamma)'V_j^{-1}Z_j\right]_{ku} \\
&\qquad \times (Z_j'V_j^{-1}Z_j)_{uk} \\
&= -\frac{1}{2}\sum_{j=1}^{J}(Z_j'V_j^{-1}Z_j)_{ku}^2 \\
&\quad + \sum_{j=1}^{J}(Z_j'V_j^{-1}Z_j)_{ku} \\
&\qquad \times \left[Z_j'V_j^{-1}(y_j - X_j\gamma)(y_j - X_j\gamma)'V_j^{-1}Z_j\right]_{ku},
\end{aligned} \tag{A.48}$$

because the matrices between brackets and parentheses in (A.48) are symmetric.

Now, from (A.33) and (A.34), we have to take expectations of the second derivatives. Therefore, from (A.4) and (A.5), the following expectations will be used:

$$\begin{aligned}
E(y_j - X_j\gamma) &= 0 \\
E(y_j - X_j\gamma)(y_j - X_j\gamma)' &= V_j.
\end{aligned}$$

From (A.36), (A.37), (A.38), and (A.39), we have

$$E\left(\frac{\partial^2 L}{\partial\gamma\,\partial\gamma'}\right) = \sum_{j=1}^{J} X_j'V_j^{-1}X_j, \tag{A.49}$$

$$E\left(\frac{\partial^2 L}{\partial\gamma\,\partial\sigma^2}\right) = 0, \tag{A.50}$$

$$E\left(\frac{\partial^2 L}{\partial\gamma\,\partial\Theta_{uv}}\right) = 0, \tag{A.51}$$

and

$$E\left(\frac{\partial^2 L}{\partial\gamma\,\partial\Theta_{uu}}\right) = 0. \tag{A.52}$$

From (A.41), (A.42), and (A.43), we have

$$\begin{aligned}
E\left(\frac{\partial^2 L}{\partial\sigma^2\,\partial\sigma^2}\right) &= -\frac{1}{2}\sum_{j=1}^{J}\mathrm{tr}V_j^{-2} + \sum_{j=1}^{J}E\left[(y_j - X_j\gamma)'V_j^{-3}(y_j - X_j\gamma)\right] \\
&= -\frac{1}{2}\sum_{j=1}^{J}\mathrm{tr}V_j^{-2} + \sum_{j=1}^{J}E\,\mathrm{tr}\left[V_j^{-3}(y_j - X_j\gamma)(y_j - X_j\gamma)'\right]
\end{aligned}$$

$$= \frac{1}{2}\sum_{j=1}^{J} \mathrm{tr} V_j^{-2}; \tag{A.53}$$

$$\begin{aligned}\frac{\partial^2 L}{\partial\sigma^2\,\partial\Theta_{kl}} &= -\sum_{j=1}^{J}(Z_j'V_j^{-2}Z_j)_{kl} \\ &\quad + \sum_{j=1}^{J} E\left\{\left[Z_j'V_j^{-1}(y_j - X_j\gamma)(y_j - X_j\gamma)'V_j^{-2}Z_j\right]_{kl}\right. \\ &\qquad \left. + \left[Z_j'V_j^{-1}(y_j - X_j\gamma)(y_j - X_j\gamma)'V_j^{-2}Z_j\right]_{lk}\right\}, \\ &= \sum_{j=1}^{J}(Z_j'V_j^{-2}Z_j)_{kl}; \end{aligned} \tag{A.54}$$

and

$$\begin{aligned} E\left(\frac{\partial^2 L}{\partial\sigma^2\,\partial\Theta_{kk}}\right) &= -\frac{1}{2}\sum_{j=1}^{J}(Z_j'V_j^{-2}Z_j)_{kk} \\ &\quad + \sum_{j=1}^{J} E\left[Z_j'V_j^{-1}(y_j - X_j\gamma)(y_j - X_j\gamma)'V_j^{-2}Z_j\right]_{kk} \\ &= \frac{1}{2}\sum_{j=1}^{J}(Z_j'V_j^{-2}Z_j)_{kk}. \end{aligned} \tag{A.55}$$

From (A.45), (A.46), and (A.48), we have

$$\begin{aligned} E\left(\frac{\partial^2 L}{\partial\Theta_{kl}\,\partial\Theta_{uv}}\right) &= -\sum_{j=1}^{J}\left[(Z_j'V_j^{-1}Z_j)_{ku}(Z_j'V_j^{-1}Z_j)_{vl}\right. \\ &\qquad \left. + (Z_j'V_j^{-1}Z_j)_{kv}(Z_j'V_j^{-1}Z_j)_{ul}\right] \\ &\quad + \sum_{j=1}^{J}\left[(Z_j'V_j^{-1}Z_j)_{ku}(Z_j'V_j^{-1}Z_j)_{vl}\right. \\ &\qquad + (Z_j'V_j^{-1}Z_j)_{kv}(Z_j'V_j^{-1}Z_j)_{ul} \\ &\qquad + (Z_j'V_j^{-1}Z_j)_{ul}(Z_j'V_j^{-1}Z_j)_{kv} \\ &\qquad \left. + (Z_j'V_j^{-1}Z_j)_{vl}(Z_j'V_j^{-1}Z_j)_{ku}\right] \\ &= \sum_{j=1}^{J}\left[(Z_j'V_j^{-1}Z_j)_{ku}(Z_j'V_j^{-1}Z_j)_{vl}\right. \\ &\qquad \left. + (Z_j'V_j^{-1}Z_j)_{kv}(Z_j'V_j^{-1}Z_j)_{ul}\right], \end{aligned} \tag{A.56}$$

and, analogously,

$$E\left(\frac{\partial^2 L}{\partial\Theta_{kl}\,\partial\Theta_{uu}}\right) = \sum_{j=1}^{J}(Z_j'V_j^{-1}Z_j)_{ku}(Z_j'V_j^{-1}Z_j)_{ul}, \tag{A.57}$$

and

$$E\left(\frac{\partial^2 L}{\partial\Theta_{kk}\,\partial\Theta_{uu}}\right) = \frac{1}{2}\sum_{j=1}^{J}(Z_j'V_j^{-1}Z_j)_{ku}^2. \tag{A.58}$$

As with the function and the gradient, computationally more efficient formulas will be derived for the covariance matrix of the estimators.

Combining (A.49), (A.50), (A.51), and (A.52) with (A.9), it is found that

$$\begin{aligned}
E\left(\frac{\partial^2 L}{\partial\gamma\,\partial\gamma'}\right) &= \sum_{j=1}^{J} X_j'\left(\sigma^{-2} I_{N_j} - \sigma^{-4} Z_j \Theta G_j^{-1} Z_j'\right) X_j \\
&= \sigma^{-2}\left(\sum_{j=1}^{J} X_j' X_j\right) - \sigma^{-4}\sum_{j=1}^{J} X_j' Z_j \Theta G_j^{-1} Z_j' X_j \\
E\left(\frac{\partial^2 L}{\partial\gamma\,\partial\sigma^2}\right) &= 0, \\
E\left(\frac{\partial^2 L}{\partial\gamma\,\partial\Theta_{uv}}\right) &= 0,
\end{aligned}$$

and

$$E\left(\frac{\partial^2 L}{\partial\gamma\,\partial\Theta_{uu}}\right) = 0.$$

Combining (A.53), (A.54), and (A.55) with (A.16) and (A.14), it is found that

$$\begin{aligned}
E\left(\frac{\partial^2 L}{\partial\sigma^2\,\partial\sigma^2}\right) &= \frac{1}{2}\sum_{j=1}^{J}\sigma^{-4}(N_j - q) + \frac{1}{2}\sum_{j=1}^{J}\sigma^{-4}\mathrm{tr}G_j^{-2} \\
&= \frac{1}{2}\sigma^{-4}(N - nq) + \frac{1}{2}\sigma^{-4}\sum_{j=1}^{J}\mathrm{tr}G_j^{-2}; \\
E\left(\frac{\partial^2 L}{\partial\sigma^2\,\partial\Theta_{kl}}\right) &= \sigma^{-4}\sum_{j=1}^{J}(G_j^{-2} Z_j' Z_j)_{kl};
\end{aligned}$$

and

$$E\left(\frac{\partial^2 L}{\partial\sigma^2\,\partial\Theta_{kk}}\right) = \frac{1}{2}\sigma^{-4}\sum_{j=1}^{J}(G_j^{-2} Z_j' Z_j)_{kk}.$$

Combining (A.56), (A.57), and (A.58) with (A.13), it is found that

$$\begin{aligned}
E\left(\frac{\partial^2 L}{\partial\Theta_{kl}\,\partial\Theta_{uv}}\right) &= \sigma^{-4}\sum_{j=1}^{J}\Big[(G_j^{-1} Z_j' Z_j)_{ku}(G_j^{-1} Z_j' Z_j)_{vl} \\
&\qquad + (G_j^{-1} Z_j' Z_j)_{kv}(G_j^{-1} Z_j' Z_j)_{ul}\Big]; \\
E\left(\frac{\partial^2 L}{\partial\Theta_{kl}\,\partial\Theta_{uu}}\right) &= \sigma^{-4}\sum_{j=1}^{J}(G_j^{-1} Z_j' Z_j)_{ku}(G_j^{-1} Z_j' Z_j)_{ul}; \\
E\left(\frac{\partial^2 L}{\partial\Theta_{kk}\,\partial\Theta_{uu}}\right) &= \frac{1}{2}\sigma^{-4}\sum_{j=1}^{J}(G_j^{-1} Z_j' Z_j)_{ku}^2.
\end{aligned}$$

These formulas are implemented in the program. Note that these expressions depend on the data only through the terms $\sum_{j=1}^{J} X_j' X_j$, $Z_j' X_j$, and $Z_j' Z_j$, which are also used for

the function and gradient (cf. section A.3), so that no additional memory is required for data storage.

Let H be the matrix defined by these expressions. Then

$$\frac{H}{N} \xrightarrow{p} \lim_{N\to\infty} \left(\frac{I(\theta)}{N}\right),$$

where $I(\theta)$ is given by equation (A.34), and θ is the parameter vector that has to be estimated. So $(H/N)^{-1}$ is a consistent estimator of the asymptotic covariance matrix of $\sqrt{N}(\hat{\theta}-\theta)$, or H^{-1} is the estimator of the covariance matrix of $\hat{\theta}$.

A.5 Reparametrization

In the formulas of the previous sections, all parameters were treated as free parameters. But, σ^2 should obviously be nonnegative, because it is a variance. Similarly, Θ should be a positive (semi-)definite matrix, because it is a covariance matrix.

To impose these restrictions, the parameters can be written in the following way:

$$\sigma^2 = (\sigma)^2 \tag{A.59}$$

$$\Theta = CC', \tag{A.60}$$

where C is a lower triangular matrix (i.e., with zero elements above the diagonal). Equation (A.59) states that σ should be the parameter used by the program, not σ^2. Equation (A.60) expresses Θ in its Cholesky decomposition, and the elements of C should be the parameters used by the program. This reparametrization may have some drawbacks (cf. Gill, Murray, & Wright, 1981, pp. 268–269), but we think that it may generally be useful for multilevel analysis. See also Longford (1987), who uses a similar reparametrization of a restricted model. Note that the reparametrization (A.60) cannot be easily used if some elements of Θ are restricted.

In order to minimize the reparametrized function, the gradient vector should be reparametrized accordingly. This is done by using the chain rule of partial derivatives: If the original parameter vector is denoted by θ, and the reparametrized parameter vector by ϕ, then

$$\frac{\partial L}{\partial \phi'} = \frac{\partial L}{\partial \theta'}\frac{\partial \theta}{\partial \phi'}. \tag{A.61}$$

Therefore, the formulas from section A.3 have to be postmultiplied by

$$\frac{\partial \theta}{\partial \phi'}.$$

The relevant formula for σ is

$$\frac{\partial \sigma^2}{\partial \sigma} = 2\sigma.$$

To form the relevant expression for C, consider the (k,l) and (k,k) elements of Θ, where

$k > l$:

$$\begin{aligned}\Theta_{kl} &= \sum_{u=1}^{q} C_{ku} C_{lu} \\ &= \sum_{u=1}^{l} C_{ku} C_{lu}; \\ \Theta_{kk} &= \sum_{u=1}^{q} C_{ku}^2 \\ &= \sum_{u=1}^{k} C_{ku}^2.\end{aligned}$$

So,

$$\begin{aligned}\frac{\partial \Theta_{kl}}{\partial C_{ku}} &= C_{lu}, && \text{if } u \leq l; \\ \frac{\partial \Theta_{kl}}{\partial C_{ku}} &= 0, && \text{if } u > l; \\ \frac{\partial \Theta_{kl}}{\partial C_{lu}} &= C_{ku}, && \text{if } u \leq l; \\ \frac{\partial \Theta_{kl}}{\partial C_{lu}} &= 0, && \text{if } u > l; \\ \frac{\partial \Theta_{kl}}{\partial C_{uv}} &= 0, && \text{if } u \neq k \text{ and } u \neq l; \\ \frac{\partial \Theta_{kk}}{\partial C_{ku}} &= 2C_{ku}, && \text{if } u \leq k; \\ \frac{\partial \Theta_{kk}}{\partial C_{ku}} &= 0, && \text{if } u > k; \\ \frac{\partial \Theta_{kk}}{\partial C_{uv}} &= 0, && \text{if } u \neq k.\end{aligned}$$

Consequently, if $u \geq v$,

$$\begin{aligned}\frac{\partial L}{\partial C_{uv}} &= \sum_{k=1}^{q} \sum_{l=1}^{k-1} \frac{\partial L}{\partial \Theta_{kl}} \frac{\partial \Theta_{kl}}{\partial C_{uv}} + \sum_{k=1}^{q} \frac{\partial L}{\partial \Theta_{kk}} \frac{\partial \Theta_{kk}}{\partial C_{uv}} \\ &= \sum_{l=v}^{u-1} \frac{\partial L}{\partial \Theta_{ul}} C_{lv} + \sum_{k=u+1}^{q} \frac{\partial L}{\partial \Theta_{ku}} C_{kv} + 2 \frac{\partial L}{\partial \Theta_{uu}} C_{uv}.\end{aligned}$$

These formulas are implemented in the program.

It is possible to transform the second derivatives in a similar way to obtain an estimator of the covariance matrix of the estimators. But, in general, the user will be more interested in the original parameters, and therefore, the estimates of the transformed parameters are retransformed to estimates of the original parameters, and the covariance matrix of section A.4 is used. This procedure is correct, because the transformation is a one to one mapping from the feasible region of the original parameters to the domain of the transformed parameters (except for some trivial equivalent solutions, such as σ and $-\sigma$, which lead to the same retransformed solution). Only when the estimates are near the boundary of the feasible region, the asymptotic covariance matrix may not be correct, but the usual statistical theory only applies to interior points, so boundary solutions are a problem in any parametrization.

Appendix B

Read.Me

MMMM MMMMM LLLL AAAAAAAA
MMMMM MMMMMM LLLL AAAAAAAAAA
MMMM M MMMMMMM LLLL AAAA AAAA
MMMM MM MMM MMMM LLLL AAAA AAAA
MMMM MMMM MMMM LLLL AAAA AAAA
MMMM MM MMMM LLLL AAAAAAAAAAAAAAAAAA
MMMM M MMMM LLLL AAAAAAAAAAAAAAAAAAAA
MMMM MMMM LLLL AAAA AAAA
MMMM MMMM LLLL AAAA AAAA
MMMM MMMM LLLL AAAA
MMMM MMMM LLLLLLLLLLLLLLLLLLLLLLLLLL AAAA
MMMM MMMM LLLLLLLLLLLLLLLLLLLLLLLLLL AAAA
AAAA
MULTILEVEL ANALYSIS FOR TWO LEVEL DATA AAAA
AAAA
VERSION 1.0b AAAA
AAAA
DEVELOPED BY AAAA
FRANK BUSING AAAA
ERIK MEIJER AAAA
RIEN VAN DER LEEDEN AAAA
AAAA
PUBLISHED BY AAAA
LEIDEN UNIVERSITY AAAA
FACULTY OF SOCIAL AND BEHAVIOURAL SCIENCES AAAA
DEPARTMENT OF PSYCHOMETRICS AND RESEARCH METHODOLOGY AAAA
WASSENAARSEWEG 52 AAAA
P.O. BOX 9555 AAAA
2300 RB LEIDEN AAAA
THE NETHERLANDS AAAA
PHONE +31 (0)71-273761 AAAA
FAX +31 (0)71-273619 AAAA

THIS FILE CONTAINS INFORMATION ABOUT THE FOLLOWING TOPICS:

- FILES ON THE MLA DISTRIBUTION DISK
- INSTALLATION NOTES
- PROGRAM'S MAIN FEATURES
- OTHER FEATURES
- INPUT AND OUTPUT
- CAPABILITY
- SYSTEM REQUIREMENTS
- CREDITS
- DOCUMENTATION
- FUTURE PLANS
- DISTRIBUTION

FILES ON THE MLA DISTRIBUTION DISK

```
  MLA.EXE      - multilevel analysis executable
  MLAE.EXE     - extended memory implementation of MLA
  READ.ME      - the file you're reading now
  EXAMPLE?.IN  - MLA input examples
  EXAMPLE?.OUT - MLA output examples
  SESAME.DAT   - sesame street data set
  RAT.DAT      - rat data set
  NELS.DAT     - nels data set
  JSP.DAT      - junior school project data set


INSTALLATION NOTES
------------------

    To install the program, simply copy all the files to the
  destination drive and/or directory and put the drive and/or directory in
  your PATH statement. Don't forget to make the PATH effective.

  For example, installing MLA on your C-drive in the directory C:\MLA

    md c:\mla           --> make the directory on your C-drive
    c:                  --> .. change to the C-drive
    cd\mla              --> .. and make \MLA the current directory
    copy a:*.*          --> copy all files from a:
    path %path%;c:\mla  --> add the directory to your path statement

  If you want the path to be effective from computer startup, change the
  path statement in your autoexec.bat. To run the program type

    MLA <inputfile> <outputfile>

  where <inputfile> should be replaced by the name of the input file and
  <outputfile> replaced by the name of the output file.


PROGRAM'S MAIN FEATURES
-----------------------

    MLA is a batch-driven statistical program that provides several types
  of estimates for a multilevel model with two levels including:

  o Summary statistics (mean, variance, standard deviation, etc.)
  o Ordinary least squares estimates
      - one-step OLS
      - two-steps OLS
      - OLS per level-2 unit
  o Full information maximum likelihood estimates including:
      - Standard errors
      - Test statistics
      - Probability values
  o Restricted maximum likelihood estimates including: (not implemented yet)
      - Standard errors
      - Test statistics
      - Probability values


OTHER FEATURES
--------------

  o Simulation analysis including:
      - Two kinds of simulation:
          - Bootstrap
          - Jackknife
      - Three different methods of bootstrap simulation:
          - Resampling from cases
          - Resampling with multivariate normal distribution
          - Resampling with residuals
      - Two different types of residual estimation for resampling:
          - raw residuals
```

        - shrunken residuals
      - Three different cases resampling schemes:
        - resample only level-1 units
        - resample only level-2 units
        - resample both level units
  o Constraints for estimates of the form: parameter = value
  o Technical settings options
  o Special output providing:
      - Input and output contents
      - Technical information
      - Raw and shrunken residuals
      - Posterior means
      - Simple diagnostics

INPUT AND OUTPUT

  The input, data, and output files are all ASCII files. The input file contains statements about the data, the model and other input requirements. The data file is a free-field formatted numbers-only ASCII file. The output file is also an ASCII file. If a file with the same name as the name of the output file already exists, it will be overwritten.
See documentation for further elaboration on these subjects.

CAPABILITY

  The program can handle up to 16 equations with 32 terms each. The table below gives the maximum values of the different input variables.

input variable	maximum	
# equations	16	
# parameters	32	(per equation)
# level-1 units	8000	(per level-2 unit)
# level-2 units	16000	
# variables	16000	
# bootstrap replications	16000	
# constrainst	64	

These limitations are the absolute maxima and can be somewhat lower depending on the amount of memory available.

SYSTEM REQUIREMENTS

  MLA will run on any IBM-PC/AT, PS/2 or compatible under MS-DOS, PC-DOS, DR-DOS or OS/2. A minimum of 256K of free RAM is necessary. MLA will also run in a DOS environment under WINDOWS or OS/2.

  The program DOES need a numeric coprocessor. However, non-coprocessor implementations are available from the authors. A coprocessor is highly recommended for extensive simulations or computations on large samples.

CREDITS

  The development of this program has been supported by a grant from SVO, project number 93713

  IBM-PC/AT and PS/2, PC-DOS and OS/2 are trademarks of International Business Machines.

  MS-DOS and MS-WINDOWS are registered trademarks of Microsoft Corporation.

  DR-DOS is a registered trademark of Digital Research.

DOCUMENTATION

From the same authors, an extensive manual was written for the MLA program. The manual contains an introduction to multilevel analysis, information about estimation procedures used in the program, a description of the input statements for MLA and many different examples. A technical appendix describes reparametrization and the minimization of the likelihood function.

FUTURE PLANS

- check on equations
- large sample summary statistics
- implementation of weights
- restricted maximum likelihood estimation
- other resampling methods
- fit measures

DISTRIBUTION

You can contact the authors by writing to the following address:

Leiden University
Faculty of Social and Behavioural Sciences
Department of Psychometrics and Research Methodology
Wassenaarseweg 52
P.O. Box 9555
2300 RB Leiden
The Netherlands
Phone +31 (0)71-273761
Fax +31 (0)71-273619

References

Aitkin, M. A., & Longford, N. T. (1986). Statistical modeling issues in school effectiveness studies. *Journal of the Royal Statistical Society A*, *149*, 1–43.

Anderson, T. W. (1958). *An introduction to multivariate statistical analysis.* New York: Wiley.

Blalock, H. M. (1984). Contextual-effects models: Theoretical and methodological issues. *Annual Review of Sociology*, *10*, 353–372.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods.* Newbury Park, CA: Sage.

Bryk, A. S., Raudenbush, S. W., Seltzer, M., & Congdon, R. T. (1988). *An introduction to HLM: Computer program and user's guide.* University of Chicago.

Busing, F. M. T. A. (1993). *Distribution characteristics of variance estimates in two-level models; A Monte Carlo study* (Tech. Rep. No. PRM 93-04). Leiden, The Netherlands: Leiden University, Department of Psychometrics and Research Methodology.

De Leeuw, J., & Kreft, I. G. G. (1986). Random coefficient models for multilevel analysis. *Journal of Educational Statistics*, *11*, 57–86.

De Leeuw, J., & Kreft, I. G. G. (1993, October). *Questioning multilevel models.* Paper presented at the Multilevel Modeling Workshop at the Rand Corporation, Santa Monica, CA.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood for incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, *39*, 1–38.

Durbin, J. (1973). *Distribution theory for tests based on the sample distribution function.* Philadelphia: SIAM.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, *7*, 1–26.

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans.* Philadelphia: SIAM.

Gill, P. E., Murray, W., & Wright, M. H. (1981). *Practical optimization.* London: Academic Press.

Glasnapp, D. R., & Poggio, J. P. (1985). *Essentials of statistical analysis for the behavioral sciences.* Columbus, OH: Charles Merrill.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika*, *73*, 43–56.

Goldstein, H. (1987). *Multilevel models in educational and social research.* London: Oxford University Press.

Goldstein, H. (1989). Restricted (unbiased) iterative generalized least squares estimation. *Biometrika*, *76*, 622–623.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, *72*, 320–340.

Kreft, I. G. G. (1994). *Are multilevel techniques necessary? An attempt at demystification.* Unpublished manuscript, California State University, School of Education, Los Angeles.

Kreft, I. G. G., & De Leeuw, J. (1991). Model based ranking of schools. *International Journal of Educational Research*, *15*, 45–59.

Kreft, I. G. G., De Leeuw, J., & Van Der Leeden, R. (1994). A review of five multilevel analysis programs: BMDP-5V, GENMOD, HLM, ML3, VARCL. *American Statistician*, *48*, 324–335.

Kreft, I. G. G., & Van Der Leeden, R. (1994). *Random coefficient linear regression models* (Tech. Rep. No. PRM-03-94). Leiden, The Netherlands: Leiden University, Department of Psychometrics and Research Methodology.

Longford, N. T. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika*, *74*, 817–827.

Longford, N. T. (1990). *VARCL. Software for variance component analysis of data with nested random effects (maximum likelihood).* Princeton, NJ: Educational Testing Service.

Longford, N. T. (1993). *Random coefficient models.* Oxford: Clarendon Press.

Maddala, G. S. (1977). *Econometrics.* Singapore: McGraw-Hill.

Magnus, J. R. (1978). Maximum likelihood estimation of the GLS model with unknown parameters in the disturbance covariance matrix. *Journal of Econometrics*, *7*, 281–312.

Magnus, J. R., & Neudecker, H. (1985). Matrix differential calculus with applications to simple, Hadamard and Kronecker products. *Journal of Mathematical Psychology*, *29*, 474–492.

Magnus, J. R., & Neudecker, H. (1988). *Matrix differential calculus with applications in statistics and econometrics.* Chichester: Wiley.

Markus, M. T. (1994). *Bootstrap confidence regions in nonlinear multivariate analysis.* Leiden: DSWO Press.

Mason, W. M., Wong, G. M., & Entwistle, B. (1983). Contextual analysis through the multilevel linear model. In S. Leinhardt (Ed.), *Sociological methodology* (pp. 72–103). San Francisco: Jossey-Bass.

Mood, A. M., Graybill, F. A., & Boes, D. C. (1974). *Introduction to the theory of statistics* (3rd ed.). Singapore: McGraw-Hill.

Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1986). *Numerical recipes: The art of scientific computing.* Cambridge, UK: Cambridge University Press.

Prosser, R., Rasbash, J., & Goldstein, H. (1991). *ML3. Software for three-level analysis. User's guide for V. 2.* London: University of London, Institute of Education.

Putter, H. (1994). *Consistency of resampling methods.* Unpublished doctoral dissertation, Leiden University, Leiden, The Netherlands.

Quenouille, M. H. (1949). Approximate tests of correlation in time series. *Journal of the Royal Statistical Society B*, *11*, 18–84.

Quenouille, M. H. (1956). Notes on bias in estimation. *Biometrika*, *43*, 353–360.

Raudenbush, S. W. (1988). Educational applications of hierarchical linear models: A review. *Journal of Educational Statistics*, *13*, 85–116.

SAS Institute. (1992). *SAS/STAT software: Changes and enhancements, Release 6.07* (SAS Technical Report No. P-229). Cary, NC: Author.

Schluchter, M. D. (1988). *BMDP5V—unbalanced repeated measures models with structured covariance matrices* (Tech. Rep. No. 86). Los Angeles: BMDP Statistical

Software.

Stephens, M. A. (1974). EDF statistics for goodness of fit. *Journal of the American Statistical Association*, *69*, 730–737.

Stevens, J. P. (1990). *Intermediate statistics: A modern approach.* Hillsdale, NJ: Lawrence Erlbaum.

Strenio, J. F., Weisberg, H. I., & Bryk, A. S. (1983). Empirical Bayes estimation of individual growth-curve parameters and their relationship to covariates. *Biometrics*, *39*, 71–86.

Tukey, J. W. (1958). Bias and confidence in not-quite large samples. *Annals of Mathematical Statistics*, *29*, 614.

Van Der Leeden, R., & Busing, F. M. T. A. (1994). *First iteration versus final IGLS/RIGLS estimates in two-level models: A Monte Carlo study with ML3* (Tech. Rep. No. PRM-02-94). Leiden, The Netherlands: Leiden University, Department of Psychometrics and Research Methodology.

Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis [with discussion]. *Annals of Statistics*, *14*, 1261–1350.